

経営戦略論で用いる回帰分析

筑波大学ビジネスサイエンス系

立本博文

tatsumoto@gssm.otsuka.tsukuba.ac.jp

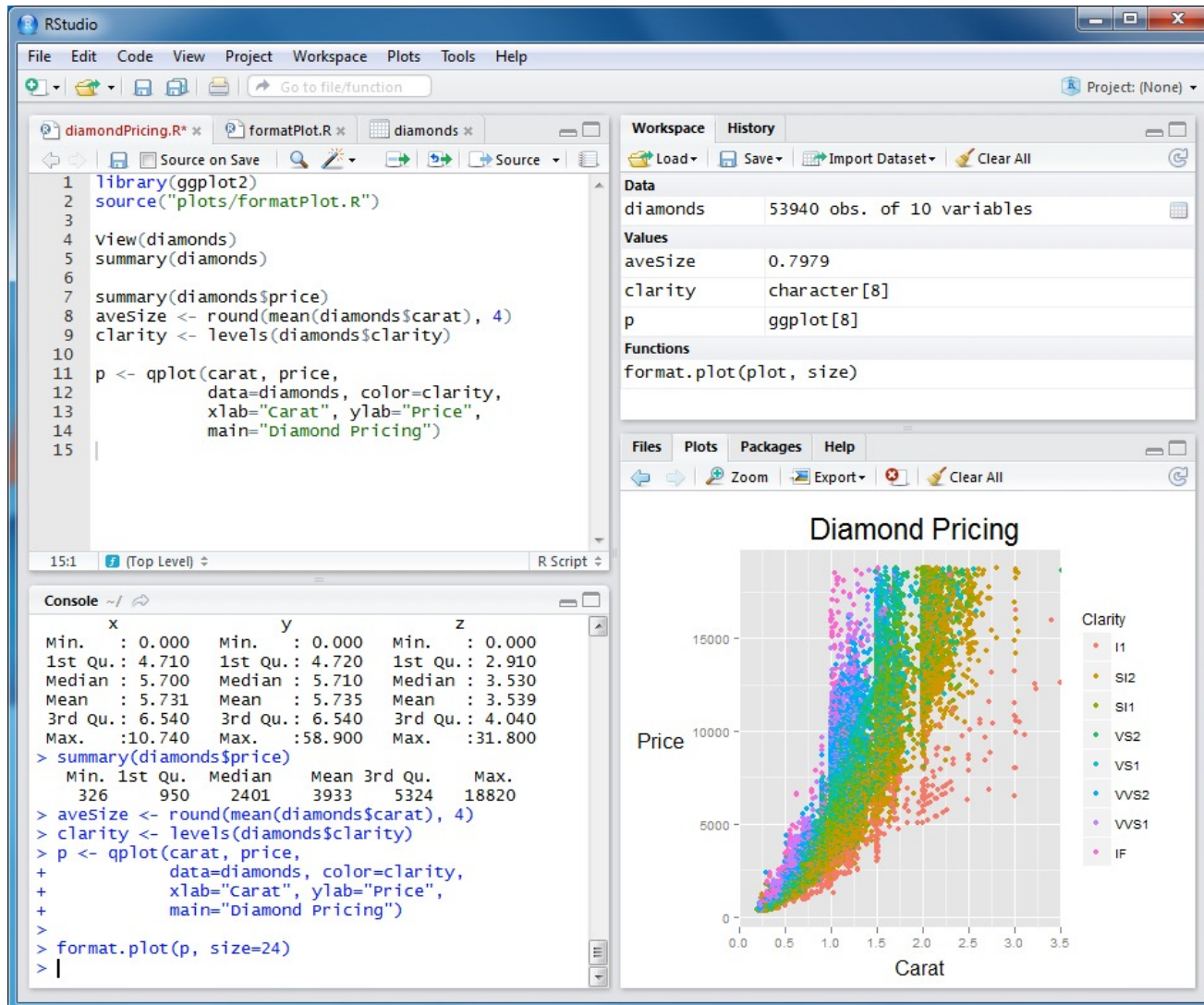
講義でやること

- 統計分析の環境を整える
- Rを使ってデータを操作する
- 経営学における回帰分析の利用
- 回帰分析を行い、分析結果の解釈をする
- 応用的な回帰モデル（交互作用モデル,媒介モデル）

統計分析の環境を整える

統計ソフトRを使えるようにする

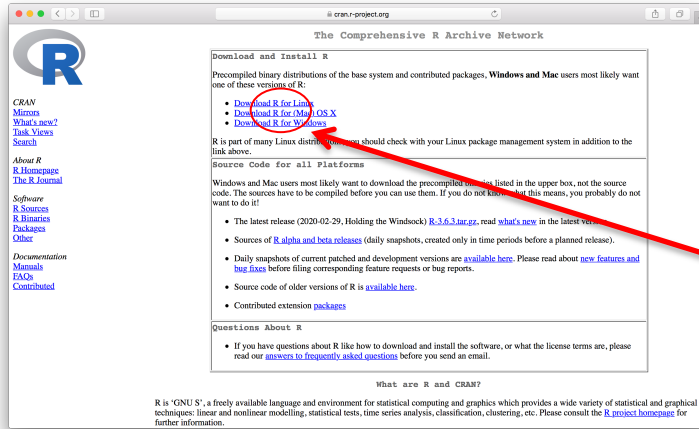
RStudio:統計ソフトRを使うための統合環境



RStudioの導入

コンピュータの操作に慣れている人は、ソフトウェアをダウンロードして自分のパソコンにインストールすることができます。

①CRANからRをダウンロードし、インストールします

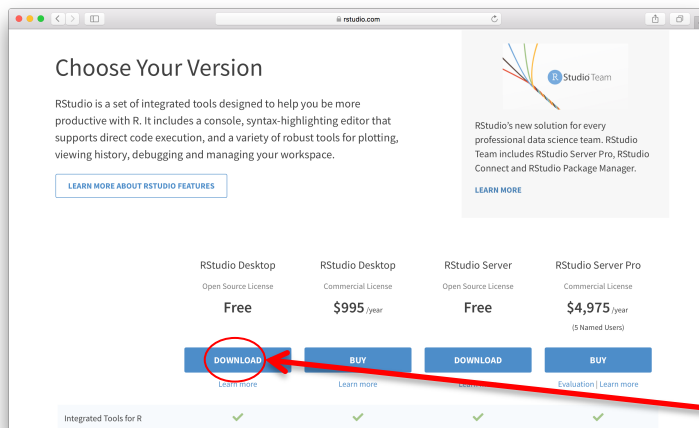


RStudioをインストールする前に、Rをインストールします

自分のOSに対応する
Rのソフトウェアをダウンロードし、
インストールする

②www.rstudio.orgからRstudioをダウンロードし、インストールします。

<https://www.rstudio.com/products/rstudio/download/>



(動画) RStudioのインストール方法

<https://www.youtube.com/watch?v=6b8pFctsBog>

RStudio Desktop Free版をダウンロードし、
インストールする

画面の使い方

The screenshot shows the RStudio IDE interface with several panes and annotations:

- Source Editor (file01.Rmd):** Contains R code. Annotations include:
 - Red circle around the Save icon: **スクリプトを保存する** (Save script).
 - Red circle around the Insert icon: **スクリプトファイルの中身** (Content of script file).
 - Red circle around the Run icon: **スクリプトを実行する** (Execute script).
- Environment Pane:** Shows the current environment. Annotation: **実行環境中の変数** (Variables in the execution environment).
- Files Pane:** Shows the project files. Annotation: **実行プロジェクト中のファイル** (Files in the execution project).
- Console/Terminal:** Shows the R prompt and output. Annotation: **コンソール** (Console).
- Output:** Displays the result of the R code execution. Annotation: **スクリプトに記録せずに直接にRを実行する** (Execute R directly without recording in the script).

The code in the Source Editor is as follows:

```
1 ---
2 title: "R Notebook"
3 output: html_notebook
4 ---
5
6 自分でメモを書き込むこともできます
7
8 ### Exercise 1 データの簡単な操作
9
10 データフレームの表示
11
12 Rではデータフレームでデータセットを表現する
13
14 ```{r}
15 #データを表示する(データフレーム名を打ち込む)
16 mtcars
17
18 ```
```

The output in the Output pane shows the first three rows of the mtcars dataset:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1

The Console shows the R prompt and the output of the mtcars command:

```
> mtcars
```

```
      mpg  cyl  disp  hp drat   wt  qsec vs  am gear carb
Mazda RX4    21.0   6  160.0  110  3.90  2.620 16.46  0   1    4    4
Mazda RX4 Wag 21.0   6  160.0  110  3.90  2.875 17.02  0   1    4    4
Datsun 710    22.8   4  108.0   93  3.85  2.320 18.61  1   1    4    1
Hornet 4 Drive 21.4   6  258.0  110  3.08  3.215 19.44  1   0    3    1
Hornet Sportabout 18.7  8  360.0  175  3.15  3.440 17.02  0   0    3    2
Valiant       18.1   6  225.0  105  2.76  3.460 20.22  1   0    3    1
Duster 360    14.3   8  360.0  245  3.21  3.570 15.84  0   0    3    4
Merc 240D     24.4   4  146.7   62  3.69  3.190 20.00  1   0    4    2
Merc 230      22.8   4  140.8   95  3.83  3.150 22.00  1   0    4    2
```

Rを使ってデータを操作する

演習

- Exercise 1 データの簡単な操作
- Exercise 2 データの抽出
- Exercise 3 平均値の差の検定
- Exercise 4 グラフのプロット
- Exercise 5 散布図行列をつくる
- Exercise 6 回帰分析を行う（単回帰）
- Exercise 7 重回帰分析を行う
- Exercise 8 交互作用モデル

Exercise 1 データの簡単な操作

1. データの表示
2. データのヘルプ(説明)の表示
3. データの次元の表示
4. データの概要の表示

Exercise 2 データの抽出

- いくつかのやり方でデータの選択・抽出を行います
- データの選択・抽出は、「条件式」と「抽出対象」を組み合わせて行います

1. 条件式の表現

2. 任意のデータを選択抽出する

3. 条件式と抽出対象を組み合わせて、選択抽出する

Exercise 3 平均値の差の検定 (t検定)

Welch Two Sample t-test

```
data: mtcars[mtcars$am == 0, "mpg"] and mtcars[mtcars$am == 1, "mpg"]
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.280194 -3.209684
sample estimates:
mean of x mean of y
17.14737 24.39231
```

注意

- 平均値の差だけを考慮するのは、第3変数の影響を考慮していない（2変数のみを考慮）
- 工場などの統制環境下のデータの分析ならば、平均値の差で判断出来ることもある
- しかし、オープン環境下（統制していない環境下）のデータでは、強いエビデンスにならない(第3変数の影響があるから)
- 第3変数の影響を調整するためには、重回帰分析を行う必要がある

Exercise 4 グラフのプロット

1. パッケージの追加
2. パッケージをロードする
3. プロット図を作成する
4. 第3変数の情報を追加する

Exercise 5 散布図行列をつくる

1. パッケージの追加
2. パッケージをロードする
3. 散布図行列を作成する
4. 適切な散布図行列を作成する

経営学における回帰分析の利用

経営学で用いる実証手法（特に因果分析）

1. 事例分析(定性分析)

メカニズムの解明に用いる

↓ 自然な分析の流れ

2. 統計分析(定量分析)

通常、実証研究(empirical study)といたら、
統計分析を指す

仮説のメカニズムが正しいとした時に、
各要因の効果の大きさを推定する

原因か否かという定性的判断と、因果的影響の強さの定量的把握とは、別の話!

経営学の研究で使われる統計手法の傾向

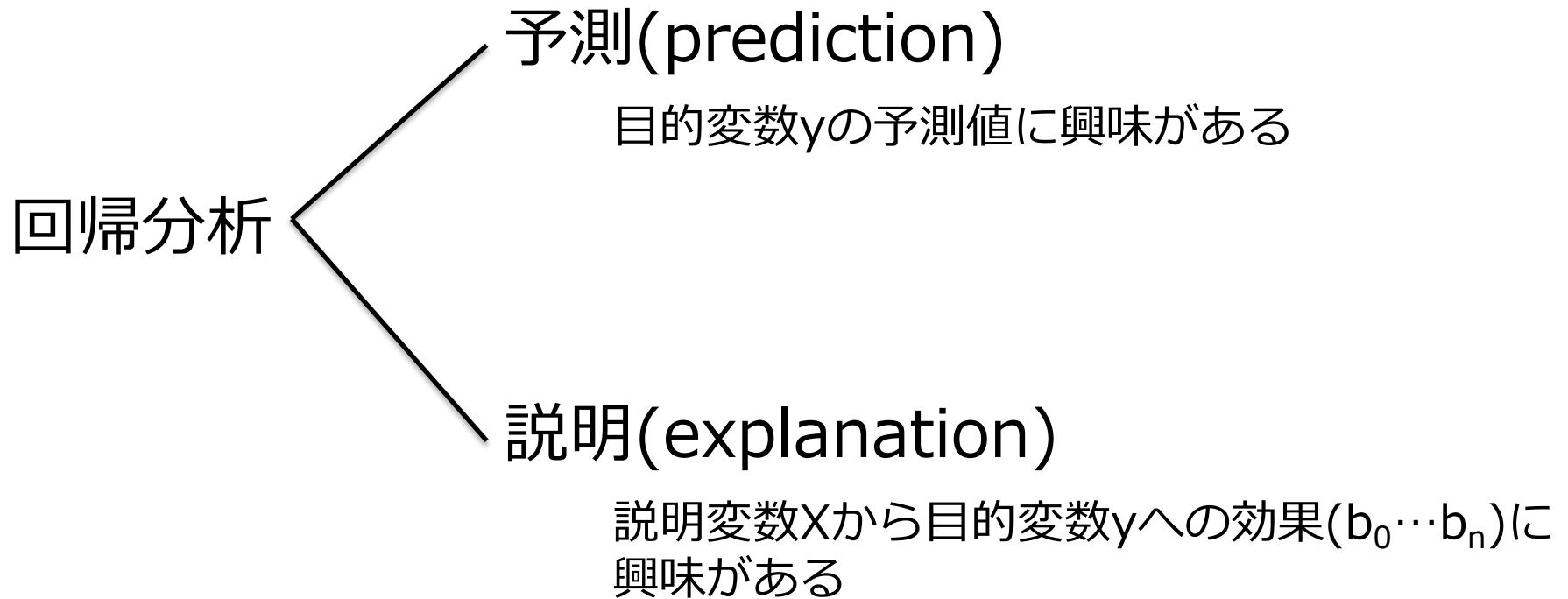
Table 1. Summary of data analytic technique usage

Data analytic technique	1980s	1990s	2000s	Correlation with year	
Frequencies	0.03	0.02	0.00	-0.229	
Nonparametric tests	0.13	0.02	0.00	-0.376†	
Correlations	0.03	0.01	0.00	-0.294	
<i>Tests of mean differences</i>	0.30	0.07	0.00	-0.704**	平均値の差の検定(t-test)は昔使われたが今は使われない
<i>t</i> -Tests	0.27	0.07	0.00	-0.622**	
Other tests of means	0.03	0.00	0.00	-0.308	
<i>General linear models</i>	0.42	0.57	0.63	0.499*	
ANOVA	0.14	0.08	0.07	-0.248	
ANCOVA	0.01	0.02	0.00	-0.085	
MANOVA	0.01	0.01	0.00	0.204	
MANCOVA	0.02	0.01	0.00	-0.227	
Simple regression	0.03	0.00	0.00	-0.358	
Multiple regression	0.19	0.34	0.39	0.567**	重回帰(multiple regression)はもっとも使われる手法
Hierarchical regression	0.02	0.11	0.17	0.681***	
<i>Longitudinal data methods</i>	0.01	0.02	0.04	0.422*	
Fixed effects models	0.00	0.00	0.00	0.292	
General time series analysis	0.00	0.01	0.00	0.052	
Pooled time series analysis	0.00	0.00	0.02	0.361†	
Variance decomposition	0.01	0.01	0.02	0.273	
<i>Explicitly dynamic methods</i>					
Event history/hazard studies	0.00	0.03	0.02	0.413†	
<i>Discrete events methods</i>	0.04	0.16	0.16	0.549**	
Discriminant analysis	0.04	0.01	0.00	-0.369†	
Financial event study	0.00	0.03	0.05	0.453*	
Logistic regression	0.00	0.12	0.11	0.757***	ロジスティック回帰は目的変数が2値(0 or 1)のときに使われる手法
<i>Methods for analysis of interdependence among firms</i>					
Network analysis	0.00	0.01	0.00	0.258	
<i>Methods explicitly accounting for firm heterogeneity</i>					
Cluster analysis	0.02	0.00	0.00	-0.083	
<i>Causal structure methods</i>	0.02	0.08	0.13	0.537**	
Path analysis	0.01	0.02	0.00	-0.209	
Simultaneous equations	0.01	0.00	0.00	-0.258	
Structural equation modeling	0.00	0.06	0.13	0.732***	構造方程式モデリング(structural equation modeling)は複数の方程式でシステムを表現するときに用いられる手法
<i>Methods to analyze decision making</i>					
Causal mapping	0.00	0.01	0.02	0.349	SEM, 共分散構造分析とも呼ぶ

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

回帰分析の目的

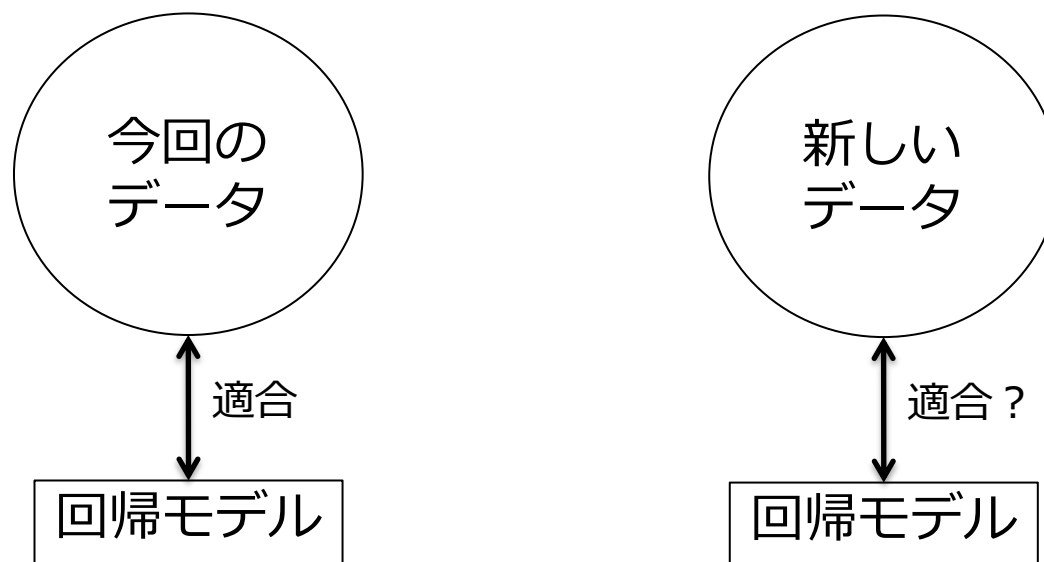
$$y = b_0 + b_1x_1 + \cdots + b_nx_n$$



(参考) 統計モデルの汎化性能

汎化性能とは・・・

母集団から新しくデータをサンプリングした時に、
当該の回帰モデルが y を予測できる性能



新しいデータで
適合度を計算

適合している
→ 汎化性能が高い

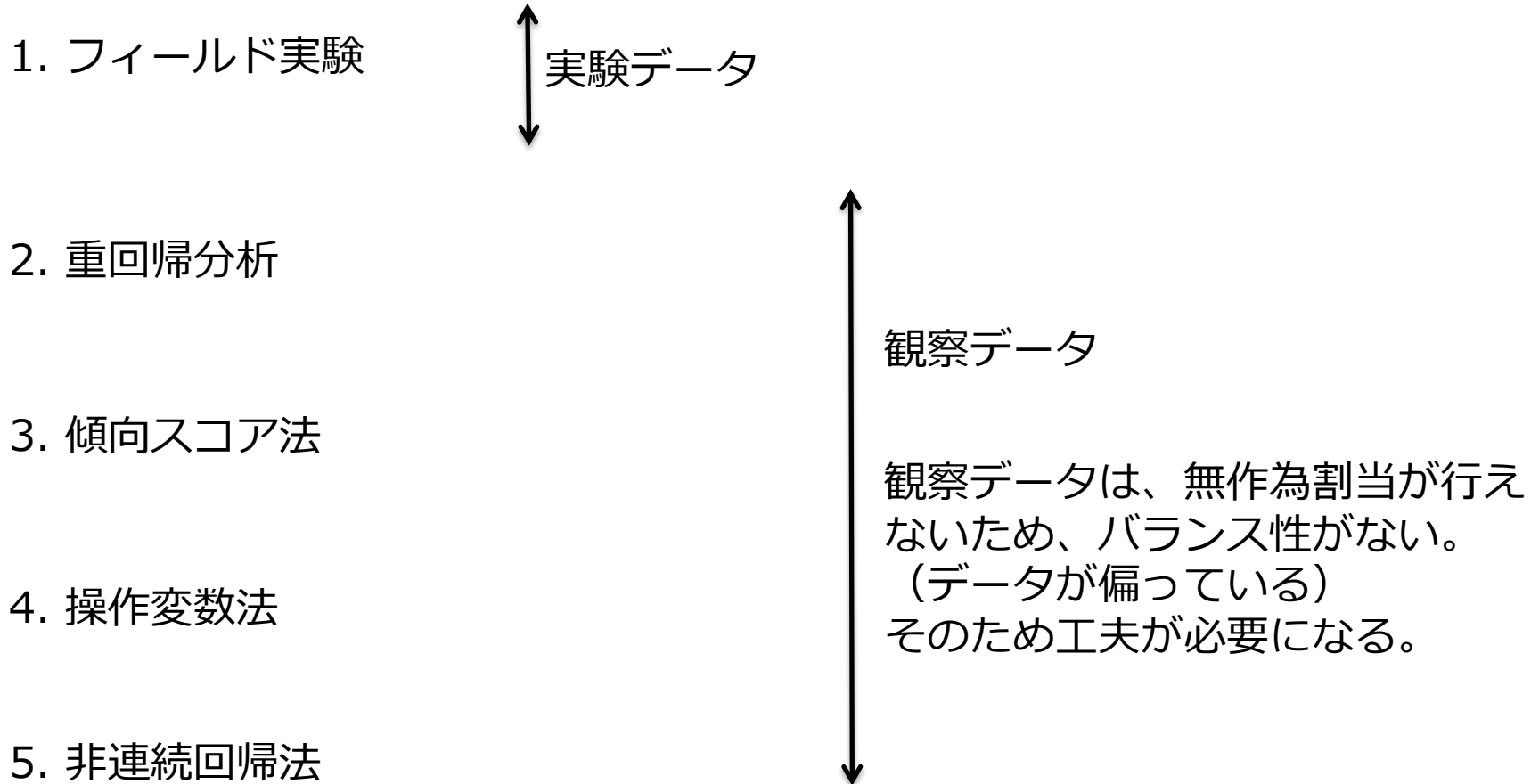
適合していない
→ 汎化性能が低い
(過剰適合・
過学習)

今回のデータに過剰に適合すると、新しいデータに対して予測力が劣る（過剰適合、過学習）。

汎化性能の指標には情報量基準や交差検証がある

回帰モデルの目的に予測が含まれる場合、汎化性能が重要になる

(参考) 因果効果量を推定する方法



Exercise 6 回帰分析をする（単回帰）

従属変数を、1つの説明変数だけで説明

$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \varepsilon \quad \varepsilon \text{に正規分布を仮定}$$

従属変数: mpg

説明変数: wt

残差(誤差): ε

回帰係数: β_1

単回帰モデルの時には回帰係数と呼ぶ

重回帰モデルの時には偏回帰係数と呼ぶ

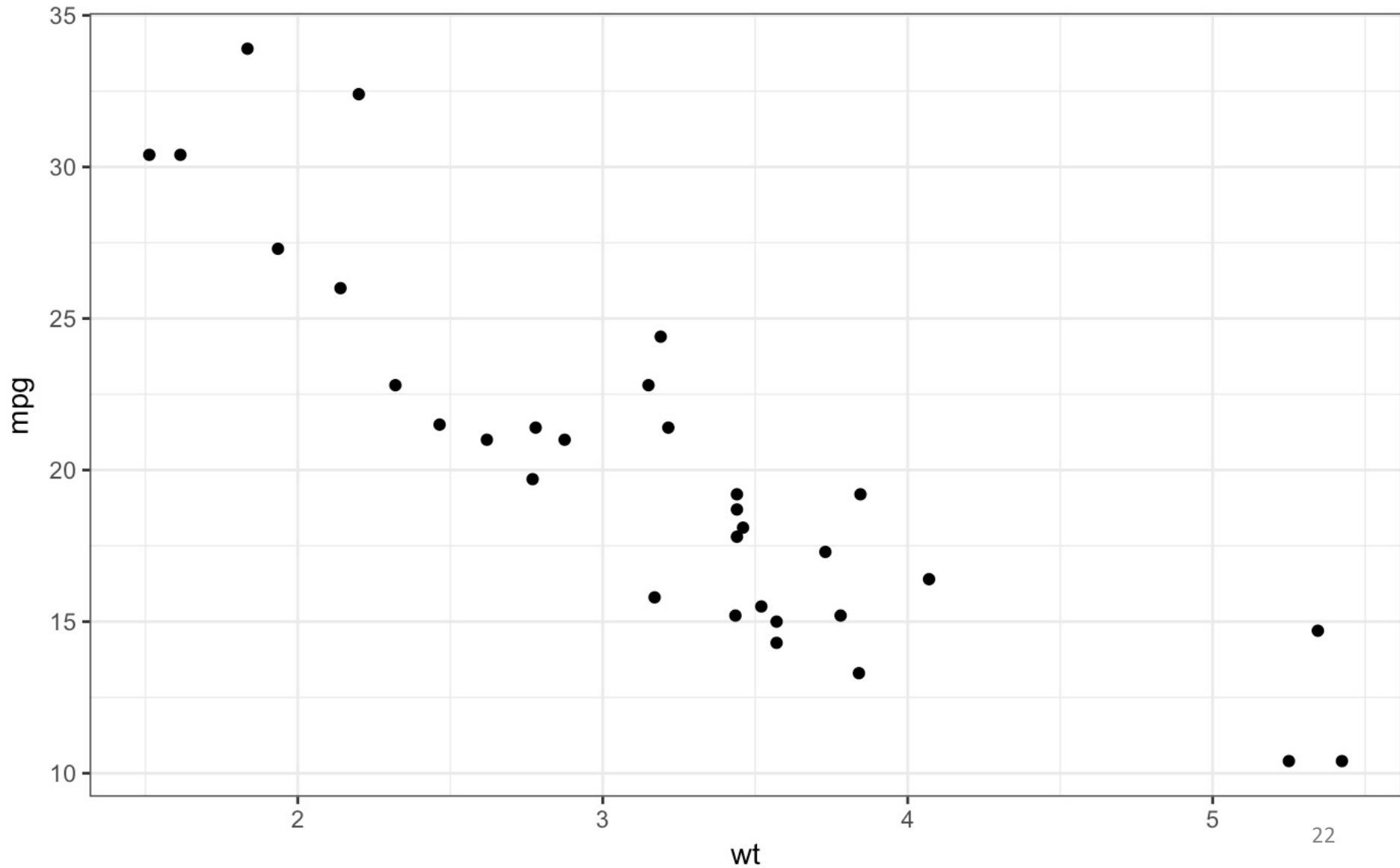
偏回帰係数の「偏」は「他の変数の影響を統制した」回帰係数という意味

データセット

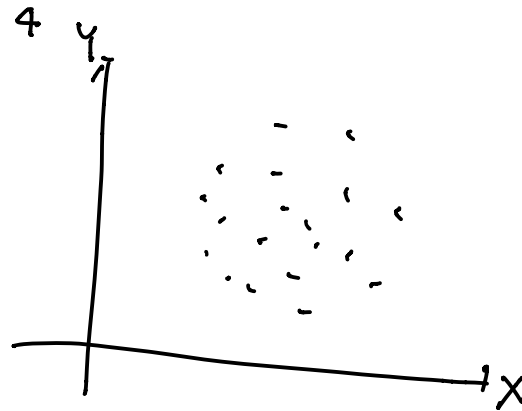
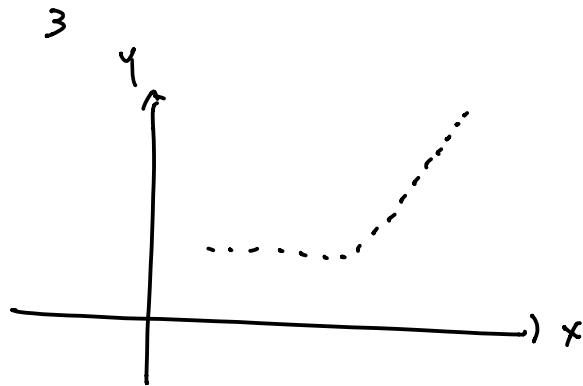
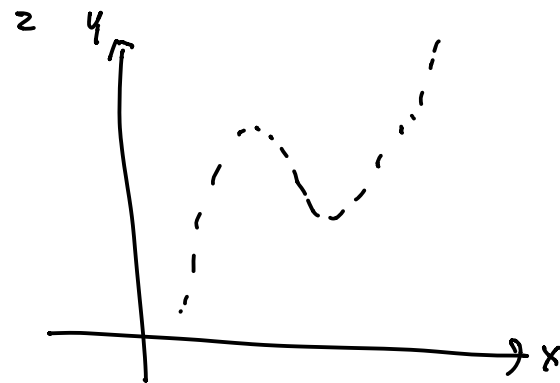
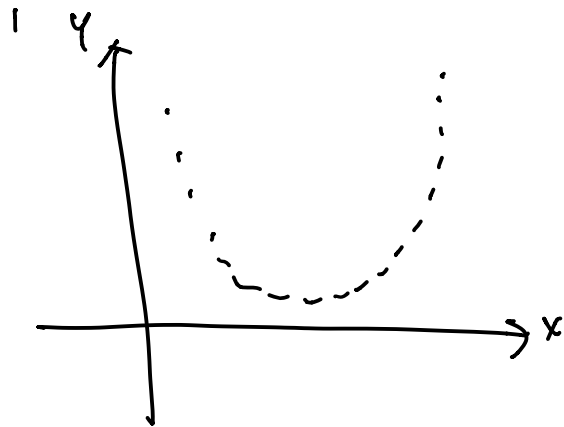
データセット	説明
mtcars	The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

変数名	説明
mpg	Miles/(US) gallon 燃費(miles/gallon)
disp	Displacement (cu.in.) 排気量(inch ³)
hp	Gross horsepower 馬力
wt	Weight (1000 lbs) 重量(1000ポンド)

Exercise 6 1.プロット図を作成する



なぜプロットをするのか



Exercise 6 2.回帰分析(単回帰)を行う

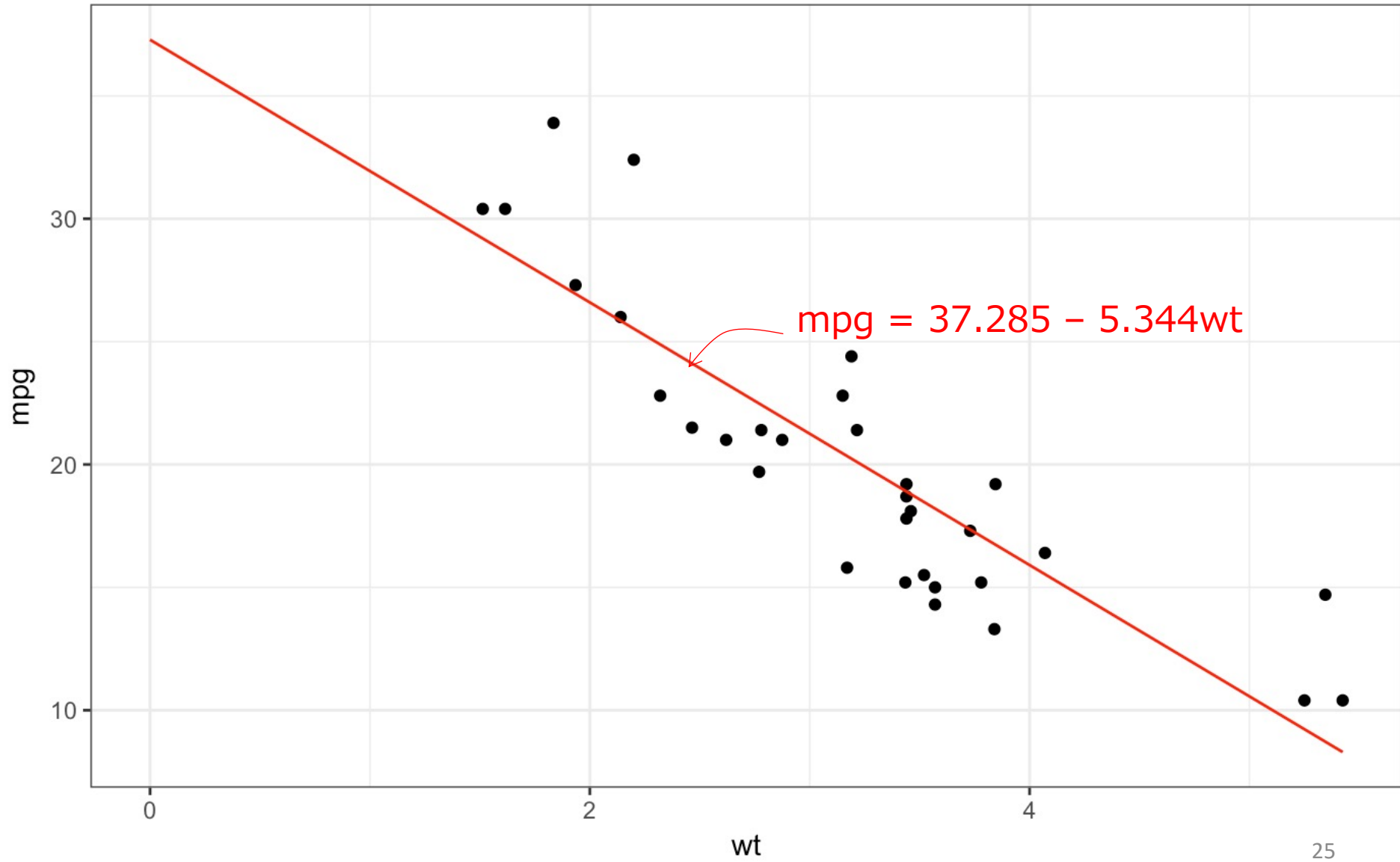
```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

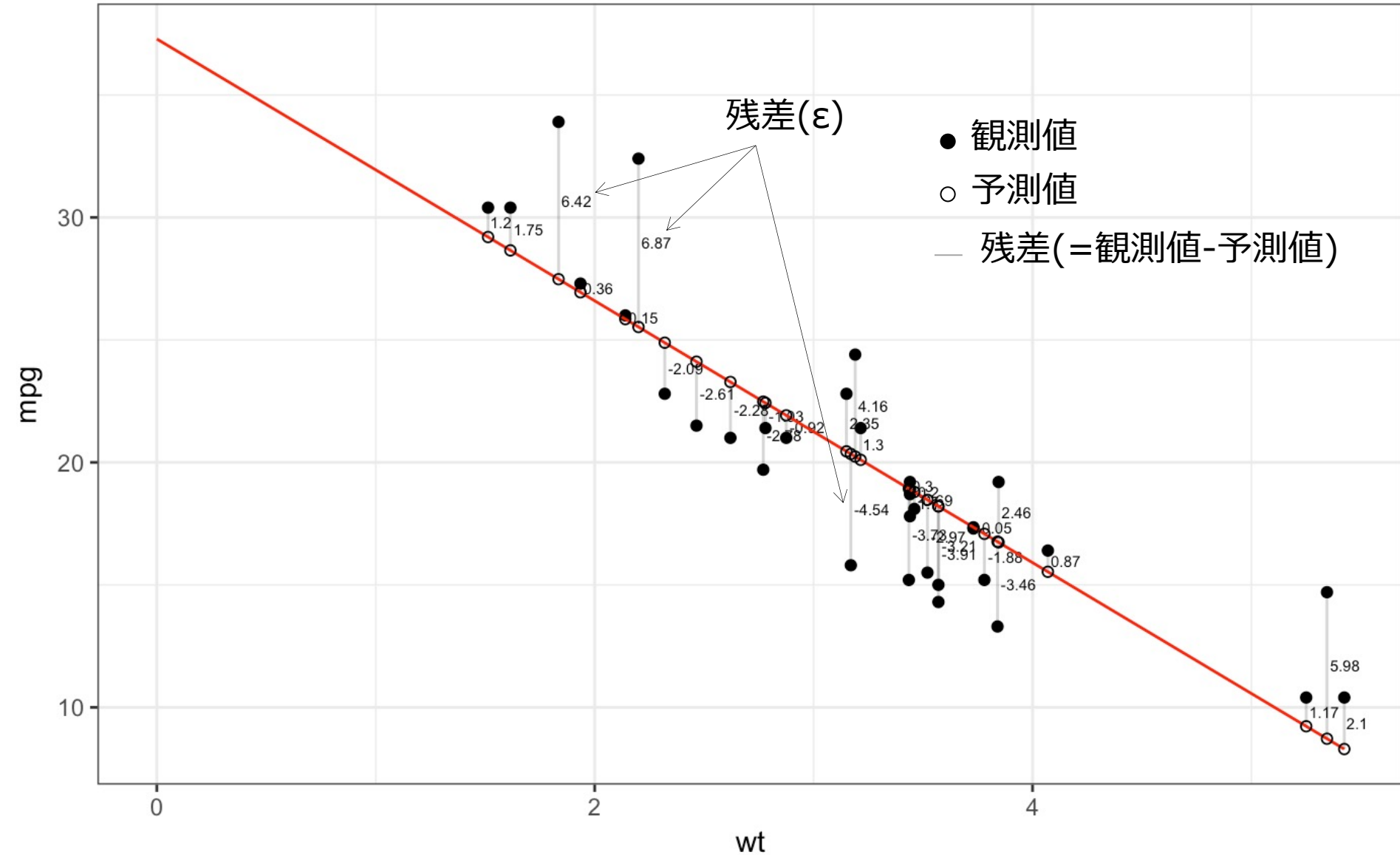
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776   19.858  < 2e-16 ***
wt          -5.3445     0.5591   -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

回歸直線



残差の表示



OLS推定：残差の二乗和を最小にするように回帰直線を推定

$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \varepsilon$$



残差 残差二乗		
ε_1	-2.283	5.210
ε_2	-0.920	0.846
ε_3	-2.086	4.351
...	...	...
ε_{30}	-2.781	7.734
ε_{31}	-3.205	10.274
ε_{32}	-1.027	1.056
和	0.000	278.322

β_0 と β_1 を調整して残差二乗の和を最小化

(最小二乗法, OLS推定)

OLS=ordinary least squares estimator

32個のデータから1本の回帰式を推定

最小二乗法



$$\text{mpg} = 37.285 - 5.344\text{wt}$$

Exercise 6 3. 回帰分析の分析結果の見方

モデル名: m1
従属変数: mpg
説明変数: wt

$$\text{mpg} = 37.285 - 5.344\text{wt}$$

Dependent variable:	
従属変数 (被説明変数) mpg m1 p値	
説明変数	wt
	-5.344***
	(0.559)
	標準誤差
Constant (切片)	37.285***
	(1.878)
赤池 情報量 基準	AIC
	166
決定係数	Observations
	観察数 (サンプルサイズ)
	32
	R ²
	0.753
自由度 調整済み 決定係数	Adjusted R ²
	0.745
F値	F Statistic
	91.375*** (df = 1; 30)
Note: *p<0.1; **p<0.05; ***p<0.01	

ミクロ的な指標
(モデルの中身の指標)

マクロ的な指標
(モデル全体に関する指標)

- ・観察数はいくつか?
→観察数が大きければ
推定結果の精度があがる
(目安: 説明変数の個数×20以上)
- ・決定係数R²はいくつか?
→R²は $0 \leq R^2 \leq 1$
十分に大きいのか?
(R²≤0.1は小さすぎる)

偏回帰係数（影響度）の評価

1. 分析対象のモデルの決定係数を確認する（適合条件）

決定係数は、目的変数の分散をモデルがどの程度説明しているのかを示す。ゆえに、決定係数は(分散)説明率とも呼ばれる。

決定係数が低すぎる場合は、偏回帰係数の解釈をしてもあまり有益でない。

2. 偏回帰係数の正負の符号を確認する(符号条件)

偏回帰係数が仮説どおりの符号であるかを確認する

偏回帰係数が仮説どおりの符号でない場合、以下を考慮する

- (i) 仮説が間違っている
- (ii) 回帰モデルが間違っている（例：コントロール変数が含まれていない）

3. 偏回帰係数が統計的有意であるのかを確認する(有意条件)

偏回帰係数が統計的有意であるかを確認する（後述）

4. 偏回帰係数が実質的に意味にある大きさかを確認する（実質科学的評価）

上記の1～3は統計的評価。これに加えて実務的に意味がある影響度かを評価

偏回帰係数が統計的有意であるのか

回帰モデルの偏回帰係数 β はサンプルデータからの推定値。
サンプリング由来の誤差を持つ。

サンプリング誤差を考慮したとしても、推定した β が意味を持つのかをテストする。
テストにはp値を用いる。

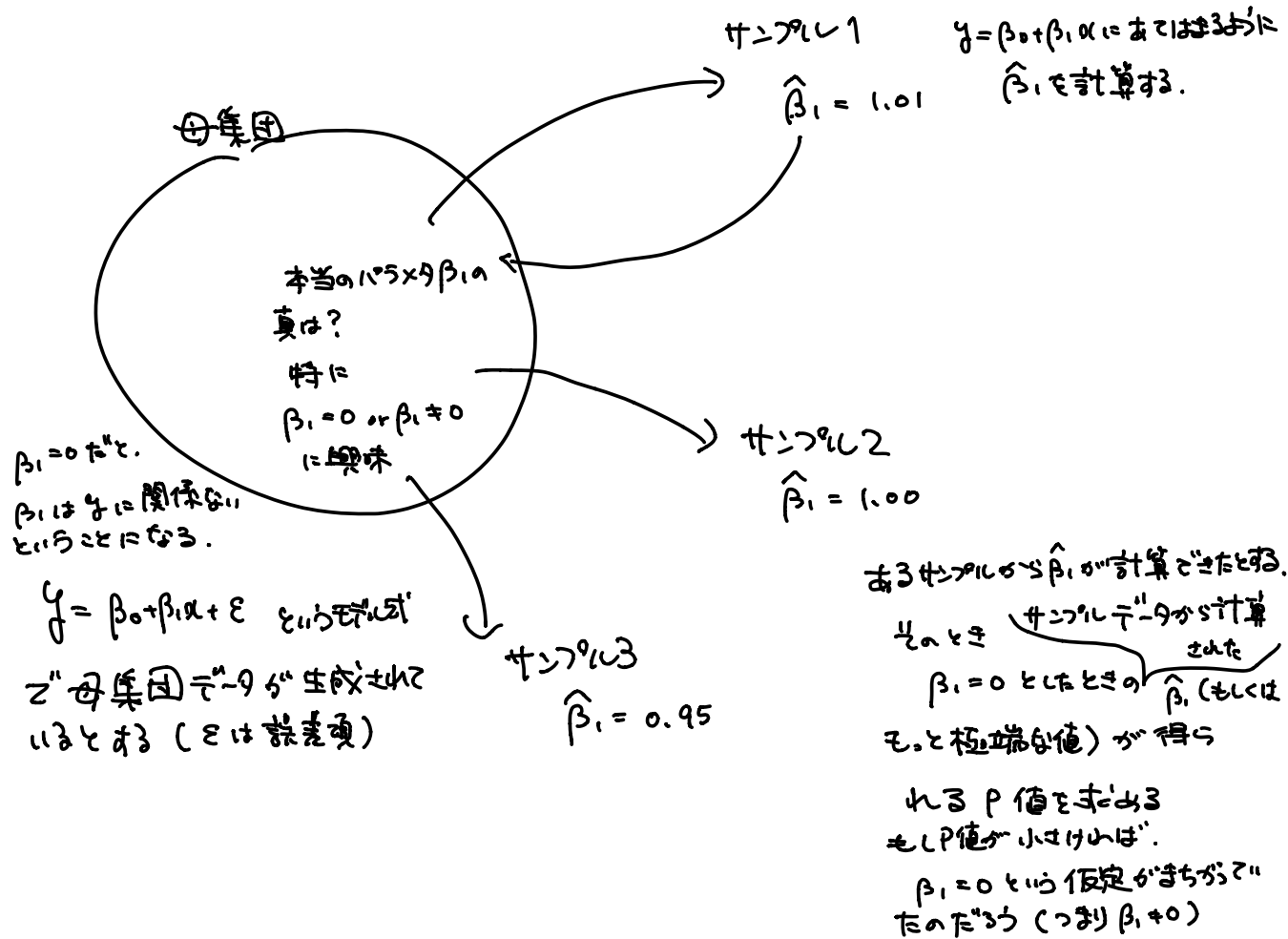
p値：
もし仮に β の真値が0とした時、 β が今回の推定値（orもっと極端な値）になる確率

→p値が小さければ $\beta=0$ である確率は小さい

p値が有意水準(慣例では5%)よりも小さければ、統計的有意とする。

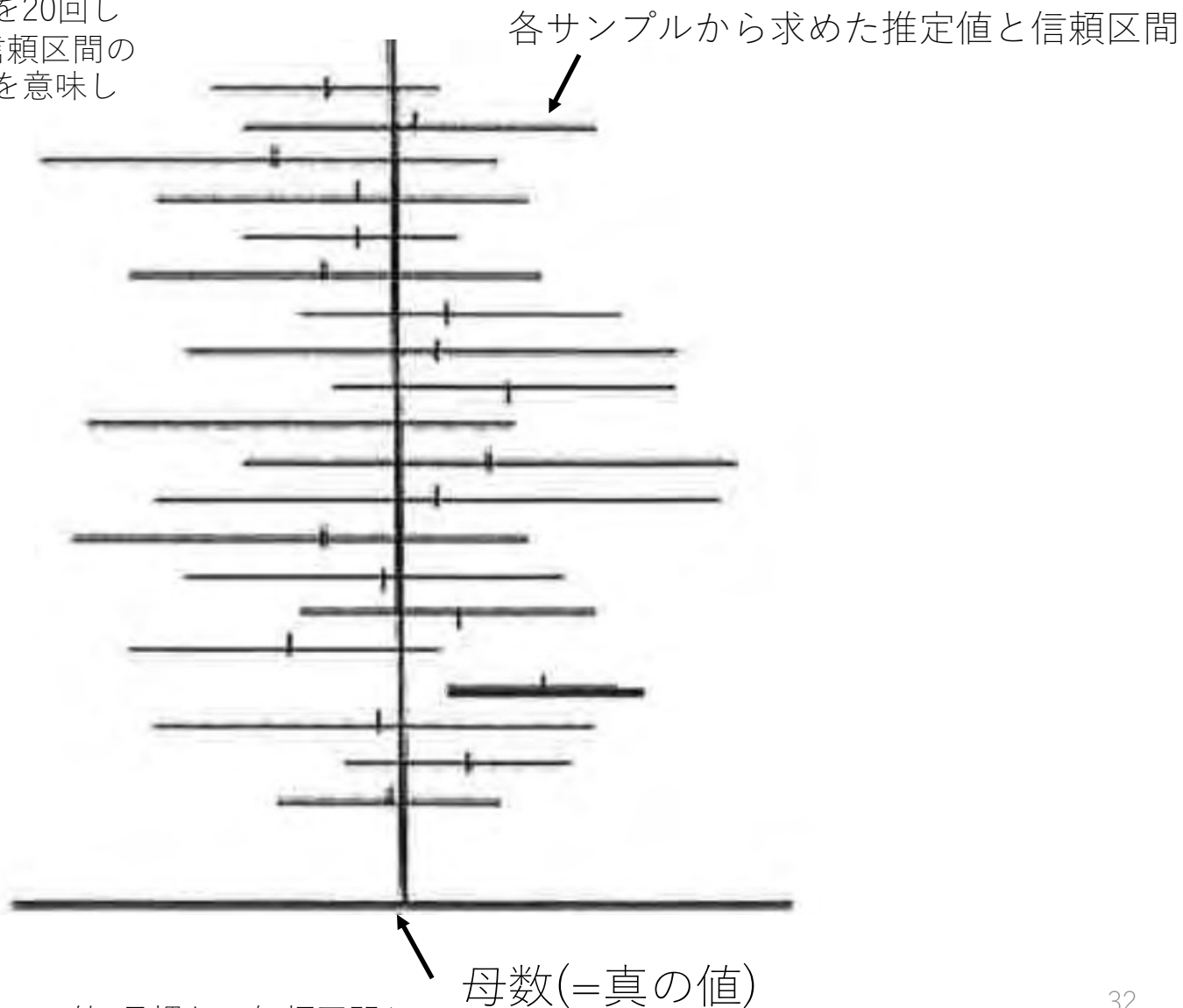
このとき「5%水準で偏回帰係数(の推定値)は統計的有意である」と表現する。

p値と帰無仮説



p値と信頼区間

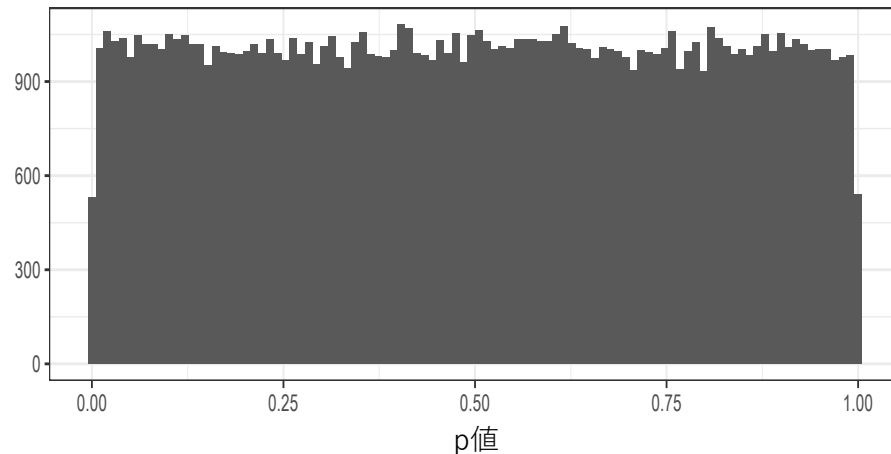
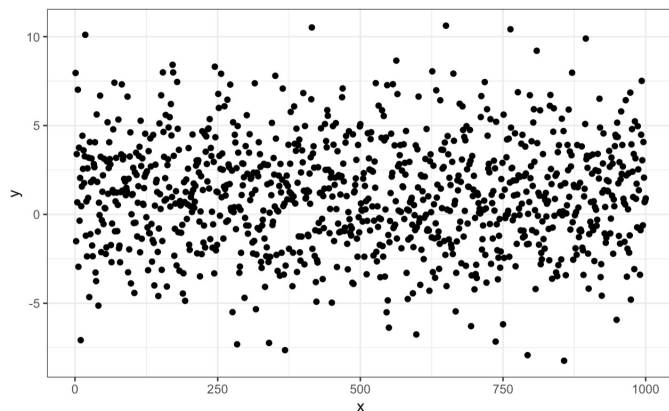
5%有意の場合、「サンプリング→推定値と信頼区間の算出」を20回したときに、1回は真の値が信頼区間の中に入らない、ということを意味している



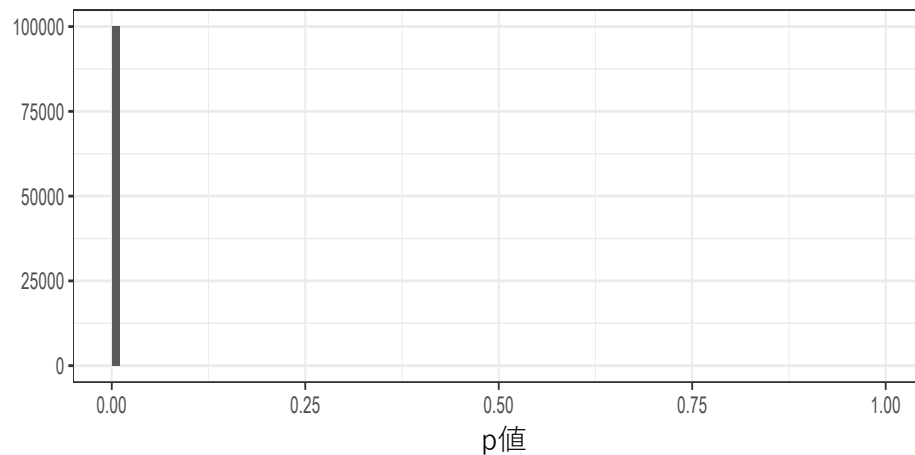
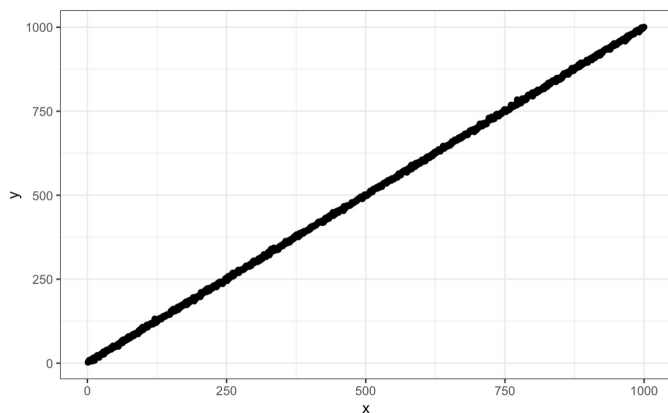
(参考) 帰無仮説のもとでのp値の分布

$y = b_0 + b_1x + \epsilon$ というモデル式で100000回データ生成し、 b_1 を推定

帰無仮説が正しい($b_1 = 0$ でデータ生成)ときのp値の分布



帰無仮説が間違っている($b_1 = 1$ でデータ生成)ときのp値の分布



Exercise 7 重回帰分析を行う

- ・ 多変量のデータセットを対象とした分析。変数間に相関がある。
- ・ 説明変数をつかって被説明変数を説明するモデル。
- ・ 説明変数の被説明変数への影響度を求めるモデル。
ただし、**影響度は他の変数の影響を統制する。**

単回帰モデルは説明変数が1つ。重回帰モデルは説明変数が複数

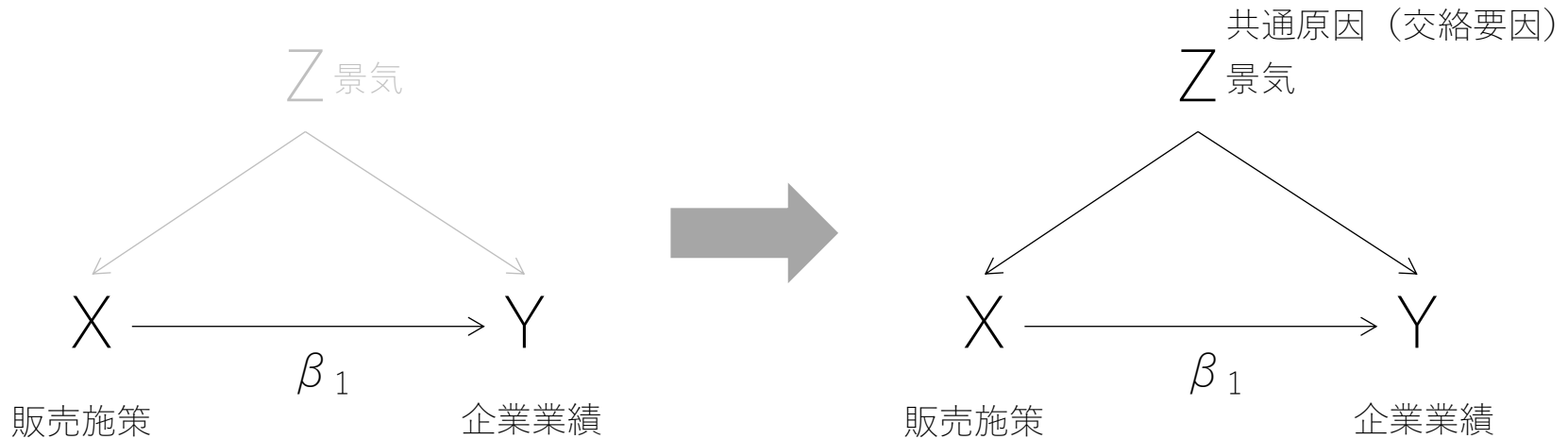
重回帰モデルの場合は以下のような式に、データがフィットするように $\beta_0, \dots, \beta_k$ を求める

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Y	:被説明変数(目的変数・従属変数)
$X_1, \dots, X_k$	:説明変数 (独立変数)
$\beta_0, \dots, \beta_k$	:偏回帰係数
ε	:残差(誤差)、正規分布に従うことを仮定
N	:観察数 (サンプルサイズ)

Exercise 7 なぜ重回帰分析をするか？

- ・ 偏回帰係数は、説明変数から被説明変数への影響度を表す。
- ・ 説明変数の被説明変数への影響度を求めるモデル。
ただし、影響度は他の変数の影響を調整する



$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 > 0$$

販売施策Xに投資したから企業業績Yが上がった？

景気Zの影響を考慮していない

$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$

$$\beta_1 > 0$$

回帰モデルに共通原因である景気Zを組み込んだ

景気Zの影響を考慮したとしても、
販売施策Xは企業業績Yにプラスの影響がある

回帰モデルに組み込んだ共通原因Zを
コントロール変数と呼ぶ

複数の回帰モデルの比較

mpg = wt ... (1)
mpg = wt + hp ... (2)

	<i>Dependent variable:</i>	
	mpg	
	(1)	(2)
wt	-5.344*** (0.559)	-3.878*** (0.633)
hp		-0.032*** (0.009)
Constant	37.285*** (1.878)	37.227*** (1.599)
Observations	32	32
R ²	0.753	0.827
Adjusted R ²	0.745	0.815
Residual Std. Error	3.046 (df = 30)	2.593 (df = 29)
F Statistic	91.375*** (df = 1; 30)	69.211*** (df = 2; 29)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01		

- ① 決定係数R², 調整済みR²は(1)→(2)で改善している(R²は十分大きい)
- ② wtの効果は、(1)では-5.344、(2)では-3.878
(2)の方が効果 (の大きさ) が小さくなっている
- ③ hpという共通原因の影響を除外したwtの影響が-3.878

(参考)階段的回帰 (Hierarchical Regression)

m1に対して、m2は良いモデルといえるだろうか？

これに答えるのが階段的回帰である

Analysis of Variance Table

Model 1: mpg ~ wt

Model 2: mpg ~ wt + hp

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)						
1	30	278.32										
2	29	195.05	1	83.274	12.381	0.001451 **						

Signif. codes:	0	***	0.001	**	0.01	*	0.05	.	0.1	'	'	1

m1に対するm2の改善を
F検定すると有意である
=m2の改善は有意である

m2はm1よりも1つ多くの変数(hp変数)を含んでいる

残差変動RSSはm2の方が小さい(m2の方がモデルの説明率が高い)
改善幅は83.27(=278.32-195.05)である

注意:現在ではAICなどの情報基準指標でモデルを比較する事が多い

(参考) HARKingは、してはだめ

HARK = Hypothesizing After the Results are Known

分析結果がわかった後に、仮説をつくってはいけない

仮説検定の手法では、危険率のぶんだけ、偶然によって統計的有意になる可能性がある。

5%有意水準なら、「20回に1回は、偶然にも、いま手にしているような分析結果を得る」という可能性がある。

(参考) p-hackingの問題

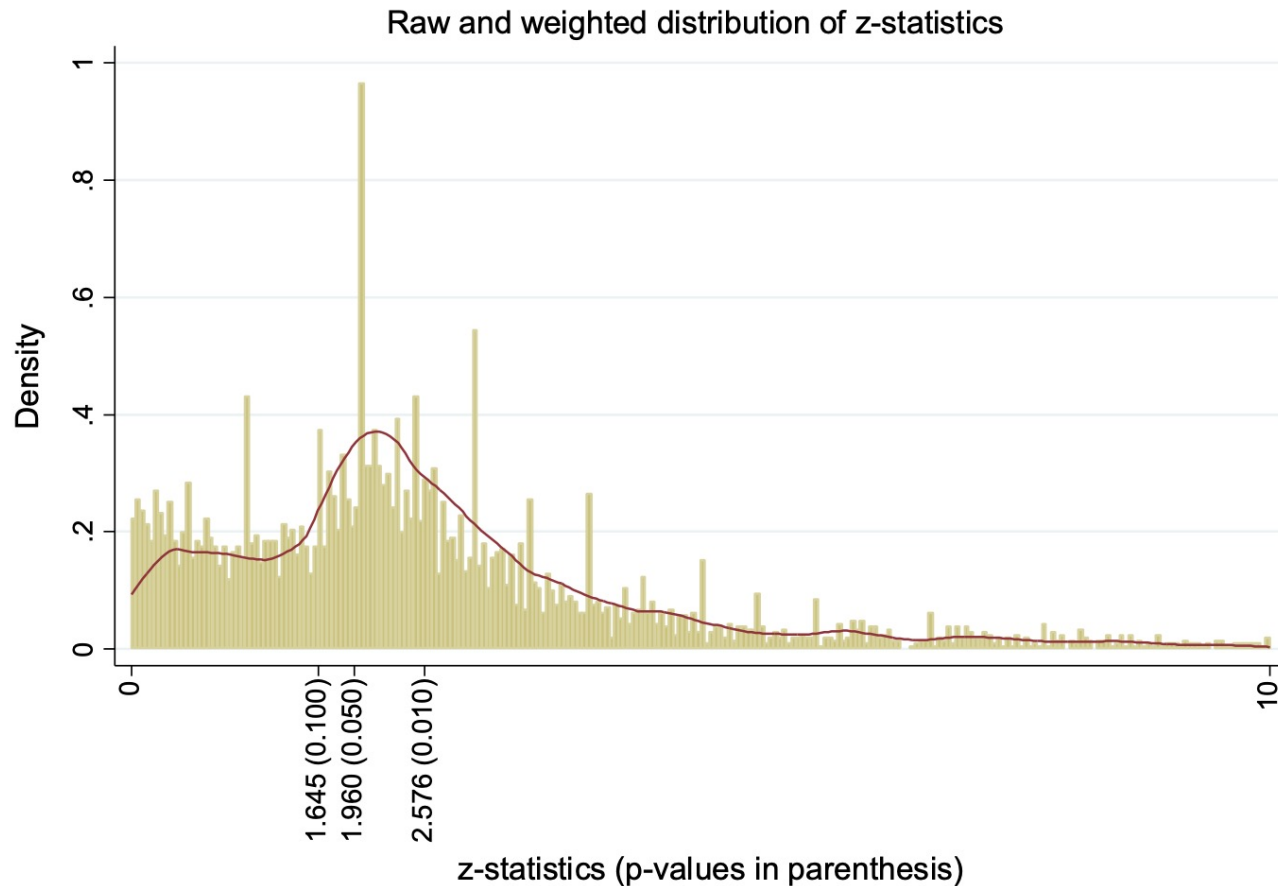


Figure 1 Camel-shaped distribution of p-values in JIBS, OrgScience and SMJ (2015 and 2016). Note The graph shows the histogram as well as the kernel density plot of the weighted distribution of z-scores in all hypotheses testing articles published in JIBS, Organization Science, and SMJ in 2015 and 2016.

課題1 平均値の差の検定

課題.Rmdファイルを見てください

課題について、データはdf_mpgを使ってください

課題1. 平均値の差の検定

mpgデータのcyl列を確認してください。

cylはシリンダ数を表し、4 5 6 8 のいずれかの値をとります。

課題1-1

cyl数が4の車種モデル群のデータの個数を答えなさい

課題1-2

cyl数が4の車両モデル群のデータの街路走行時燃費(cty)の平均値と標準偏差を答えなさい

課題1-3

cyl数が4の車種モデル群と、cyl数が8の車種モデル群で、街路走行時燃費(cty)の平均に違いがあるかを、t検定を用いて調べて、報告しなさい。

課題2. 回帰分析

課題.Rmdファイルを見てください

課題について、データはdf_mpgを使ってください

以下の2つのモデル式に対応する回帰モデルM1,M2を考えます

排気量が多いほど燃費が悪いのではないか、との仮説H1の下にモデル1を考えました。

モデルM1:

$\text{cty} = \text{displ}$

また、シリンダ数が多いほど、燃費が悪いのではないか、との仮説H2の下に、次のモデルM2を考えました

モデルM2:

$\text{cty} = \text{cyl}$

さらに、これらの仮説H1とH2を総合して、モデルM3を考えました

モデルM3:

$\text{cty} = \text{displ} + \text{cyl}$

課題2-1

街路走行時燃費(cty)、排気量(displ)、シリンダ数(cyl)の3つの変数からなる散布図行列を作成してください

課題2-2

モデルM1,M2,M3の回帰分析の推定値をまとめた回帰テーブルを作成しなさい

課題2-3

課題2-2で作成した回帰テーブルに記載されたM3の結果について、以下の点について報告しなさい。

- ・モデルの適合条件はどうですか？
- ・偏回帰係数の符号条件はどうですか？
- ・偏回帰係数の有意条件はどうですか？

統制変数の選び方

（参考）統制変数の選び方 シンプソン・パラドックス

統制変数（コントロール変数）に
よって正しく統制ができない現象

(参考) シンプソン・パラドックス

「適切な統制をしないと、 $X \rightarrow Y$ の効果が正しく推定できない」という問題を統計学ではシンプソン・パラドックスと呼んでいる

シンプソン・パラドックスの例として、UCB admissionデータが知られている

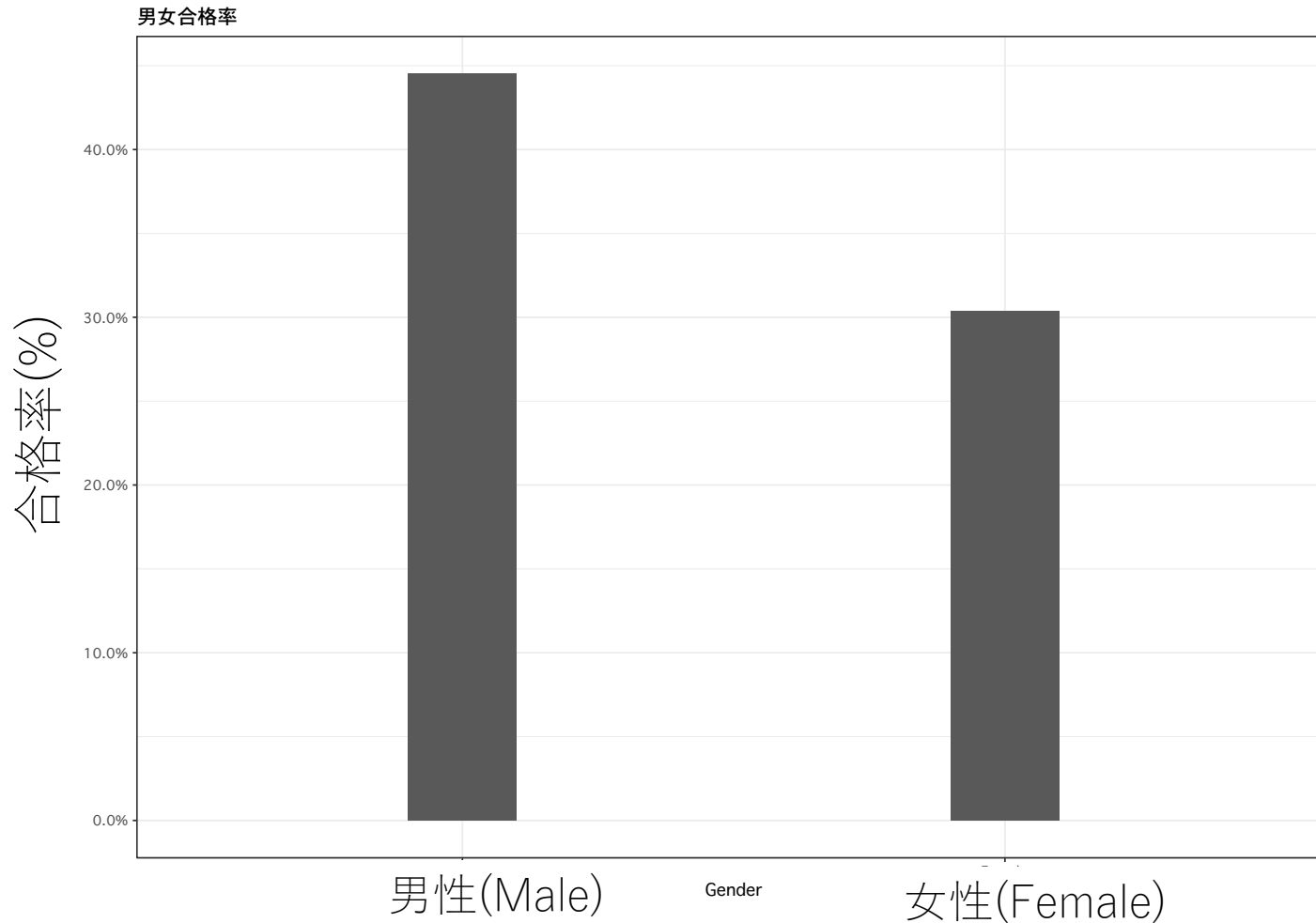
UCB Admissionは、UCB（カリフォルニア州立大学バークレイ校）の入学試験のデータである

同校では、女性受験者の合格率が低いことが、批判されていた

しかし、関係者には、性差別の意図は全く無く、当惑していた

(参考) UCB admission

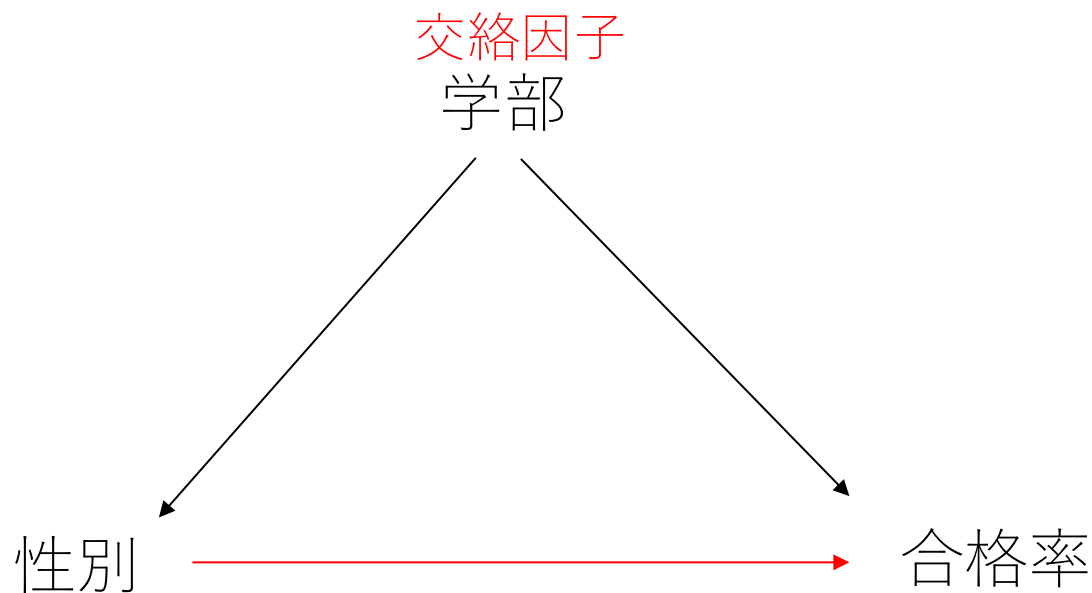
Simpson's Paradox: UC Berkeley 1973 Admissions



男性のほうが、女性よりも、入試の合格率が高い。差別ではないのか？

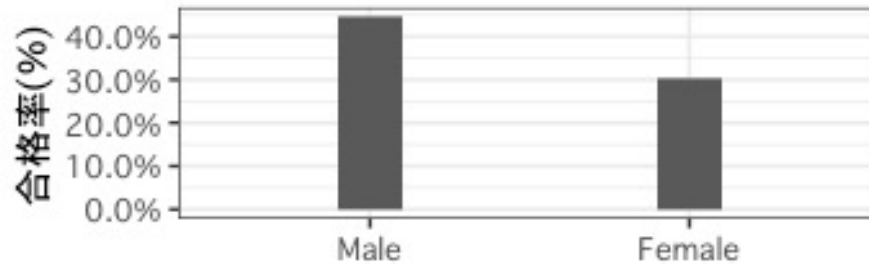
(参考) UCB Admission

学部が、合格率に影響しているのではないかな？



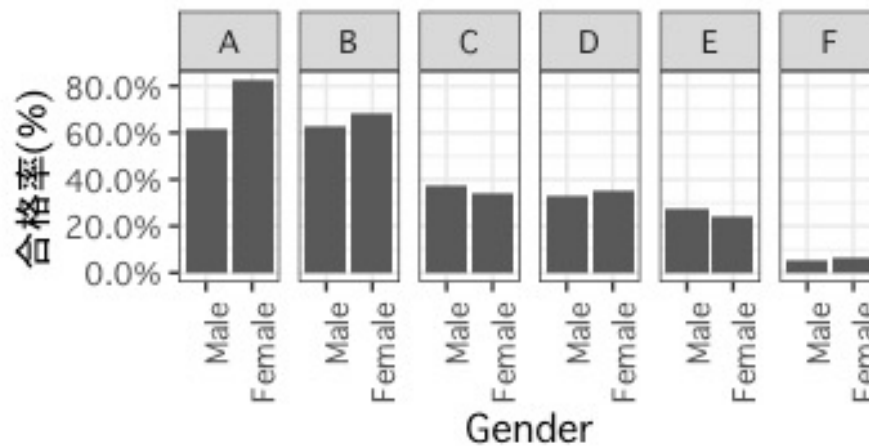
(参考) UCB admission

男女合格率



統制なしの合格率でみると、
男性の合格率が高い

学部毎合格率



学部で統制してみると、
性別による合格率の違いは
小さくなる

(参考) UCB Admissionのロジスティック回帰の例

	<i>Dependent variable:</i>	
	Admit	
	(1)	(2)
GenderMale	0.610*** (0.064)	-0.100 (0.081)
DeptB		-0.043 (0.110)
DeptC		-1.263*** (0.107)
DeptD		-1.295*** (0.106)
DeptE		-1.739*** (0.126)
DeptF		-3.306*** (0.170)
Constant	-0.830*** (0.051)	0.682*** (0.099)
Observations	24	24
Log Likelihood	-2,975.446	-2,593.744
Akaike Inf. Crit.	5,954.891	5,201.488
Note:	*p<0.1; **p<0.05; ***p<0.01	

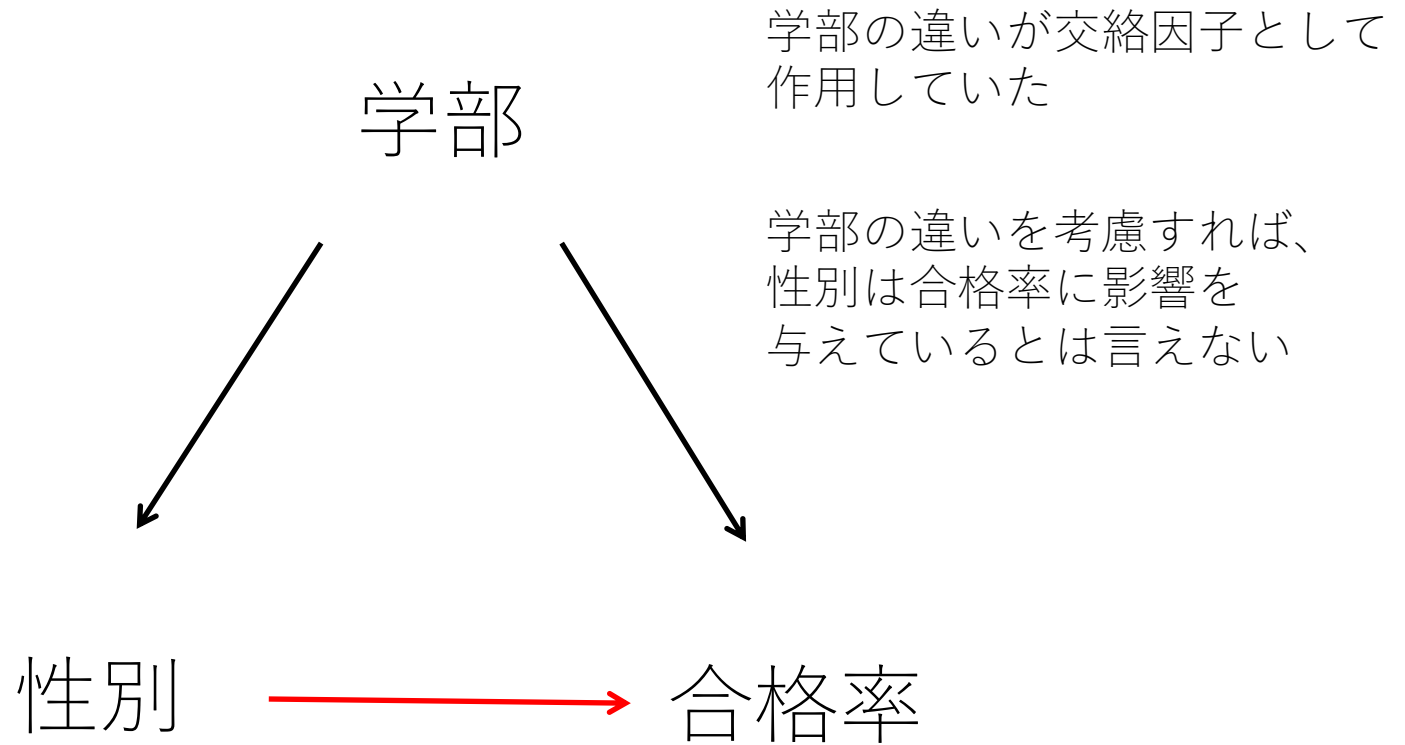
$$\exp(0.610) \div 1.84$$

性別=男性だと、
そうでない場合(=女性)に比べて
合格オッズが1.84倍ほど高くなる

しかしながら、
学部(Dept)の違いを考慮すると
性別による合格オッズの違いは
ほとんどない(=-0.100)

統計的に有意でもない
=性別で違いがあるとは言えない

(参考) UCB admission

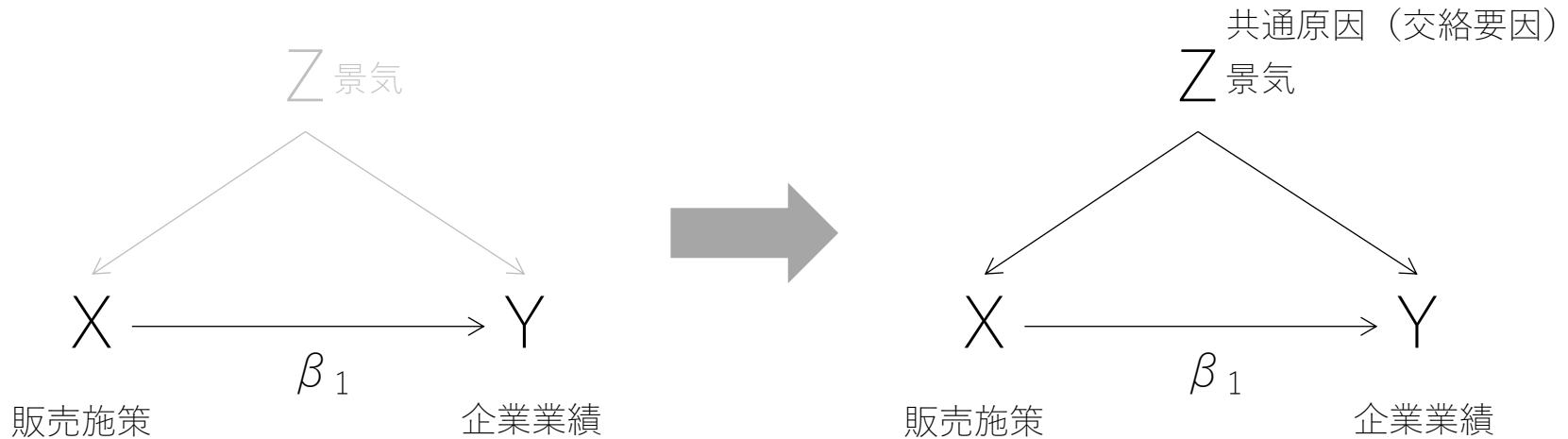


統制変数の選び方

回帰分析で適切に統制をする
交絡因子の見つけ方
バックドア基準

なぜ重回帰分析をするか？

- ・ 偏回帰係数は、説明変数から被説明変数への影響度を表す。
- ・ 説明変数の被説明変数への影響度を求めるモデル。
ただし、影響度は他の変数の影響を調整する。



$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 > 0$$

販売施策Xに投資したから企業業績Yが上がった？

景気Zの影響を考慮していない

$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$

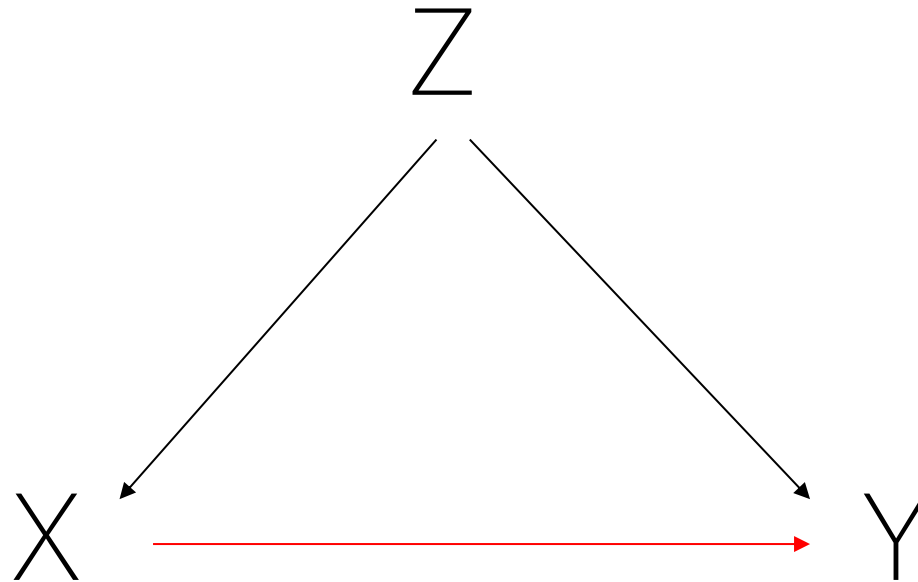
$$\beta_1 > 0$$

回帰モデルに共通原因である景気Zを組み込んだ

景気Zの影響を考慮したとしても、
販売施策Xは企業業績Yにプラスの影響がある

回帰モデルに組み込んだ共通原因Zを
コントロール変数 (統制変数) と呼ぶ

変数Zとは何なのか？



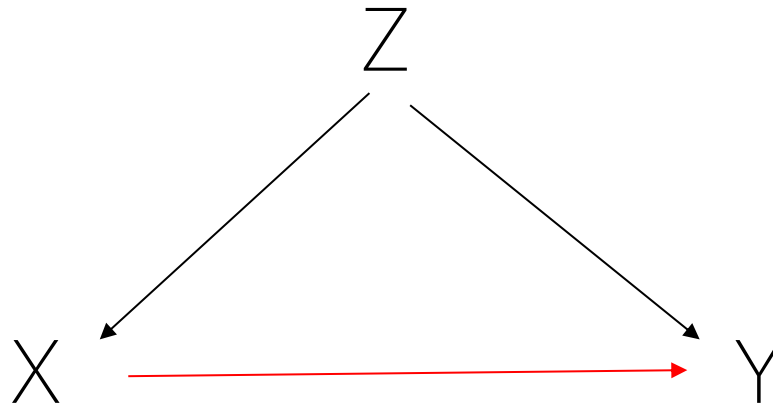
Zは、XとYの共通原因（交絡変数）

回帰分析では、XのYへの効果を推定する時には、
共通原因Zを統制変数としてモデルに含めて推定せよ、
としている

変数Zとは何なのか？

変数Zは、**交絡変数(confounder)**として統計学では知られていた

confound：狼狽させる。当惑させる。



3変数間に上記のような関係がある時に、Zで統制しないと $X \rightarrow Y$ の効果が正しく推定されない

適切な統制する = 重回帰モデルの中に**交絡変数Zを統制変数として含める**

回帰分析による交絡変数Zの影響の調整

交絡変数が存在した場合、回帰分析で統制変数として取り込んで処理すれば、正しい効果が推定できる

しかし、どのような変数が交絡変数で、どのような場合に回帰分析に取り込めばよいのか不明である

DAGを書き、back-door基準で交絡変数の選択をすればよい

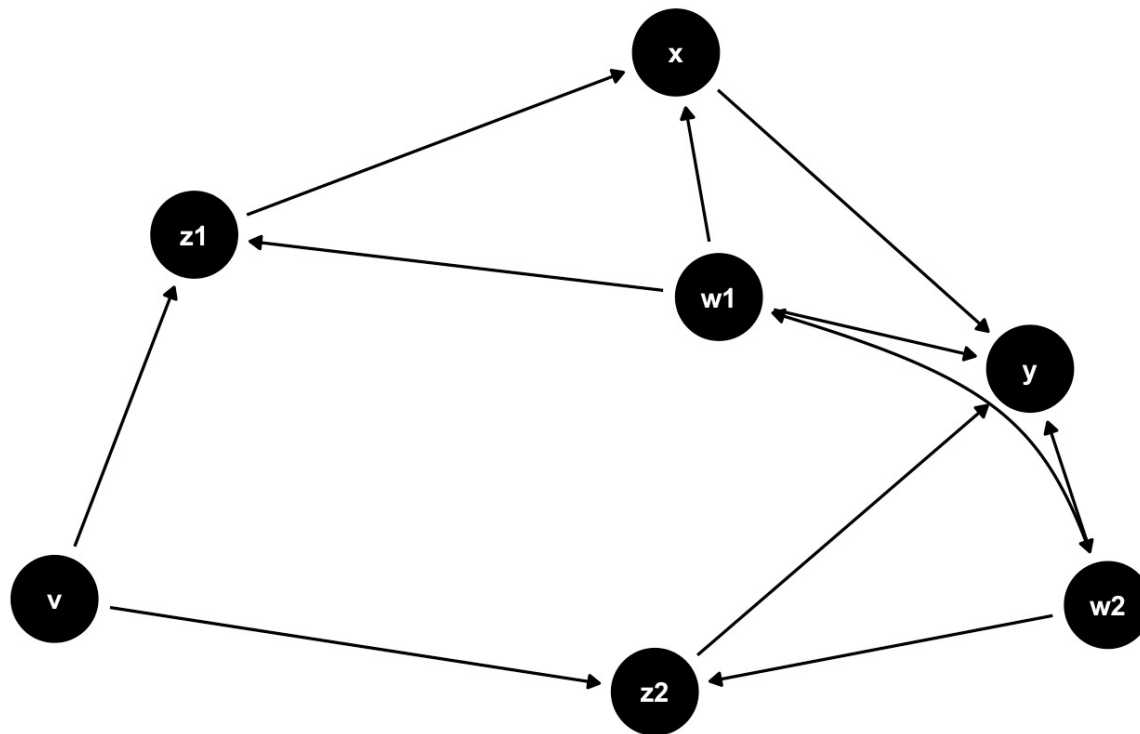
Pearl, J.(2000) Causality: Models, Reasoning and Inference

DAG: Directed Asynchronous Graph (有向非巡回グラフ)

$x, y, v - z2$ の7つの変数で表現できるシステム。 $x \rightarrow y$ の効果を知りたい

$y = b_0 + b_1x + \dots$ という回帰モデルを推定。
…に、いずれの変数を選択すればいいのか?

x : 仮説対象の説明変数
 y : 目的変数 (被説明変数)

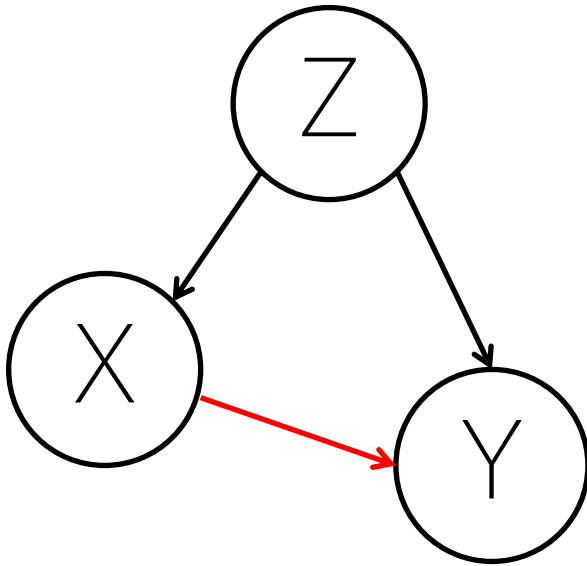


変数(共変量)は $v \sim z2$ まで5つある。どれが調整しなければならない交絡変数なのか?

バックドア基準(back-door criteria)

3変数(X,Y,Z)の場合を考える。XからYの因果効果の大きさを知りたい(→印)
その時、X→Yの効果を正しく表す b_1 を得るためには、
以下のように第3変数Zを回帰モデルに加える(もしくは、加えない)。

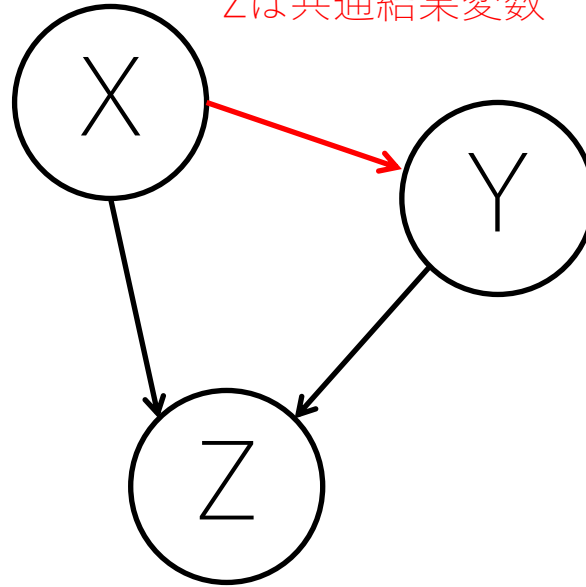
(1) Zが分岐点
Zは交絡変数



Zを回帰モデルに**加える**

$$Y = b_0 + \mathbf{b_1}X + b_2Z$$

(2) Zが合流点
Zは共通結果変数

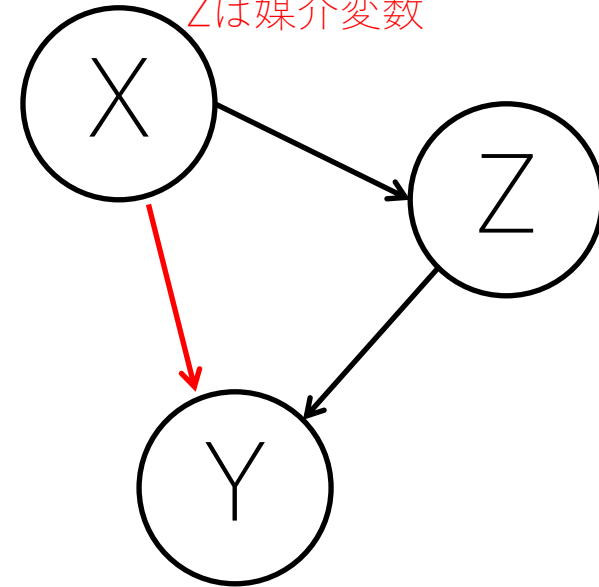


Zを回帰モデルに**加えては
ならない**

$$Y = b_0 + \mathbf{b_1}X$$

Zを含めるとX→Yの
偽の関連性(疑似相関)が
発生する

(3) Zが中間点
Zは媒介変数



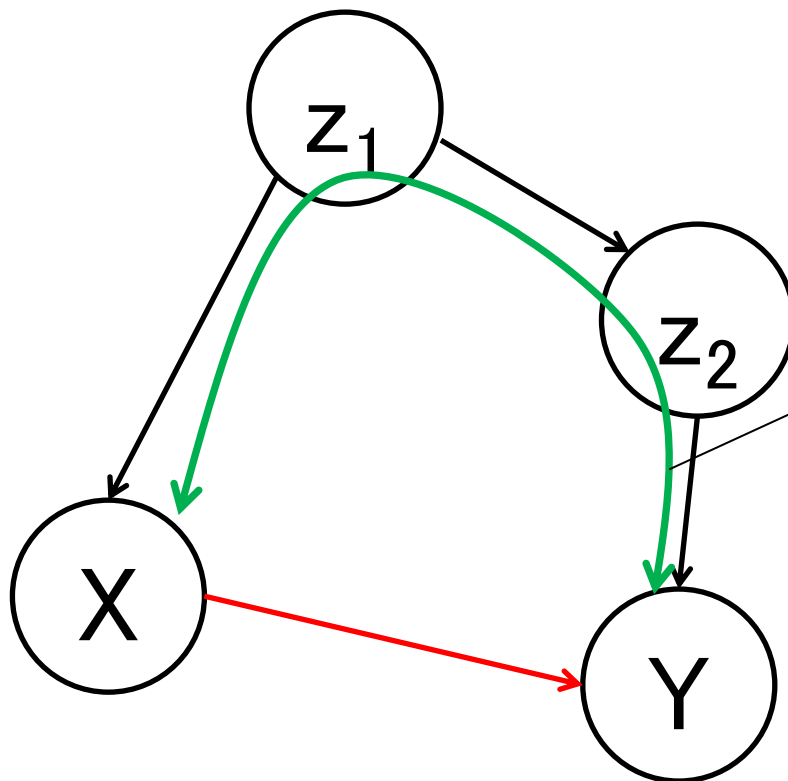
総合効果を知りたい時はZを回帰モデルに
加えてはならない
直接効果を知りたい時はZを回帰モデルに
加える

(総合効果) $Y = b_0 + \mathbf{b_1}X$

(直接効果) $Y = b_0 + \mathbf{b_1}X + b_2Z$

4変数以上の場合のバックドア基準

4変数以上の場合、道(path)で考える



このpathを構成する
いずれかの変数が
回帰モデルに含まれれば
 $X \rightarrow Y$ の因果効果は正しく
推定できる

$$(1) Y = b_0 + b_1 X + b_2 Z_1$$

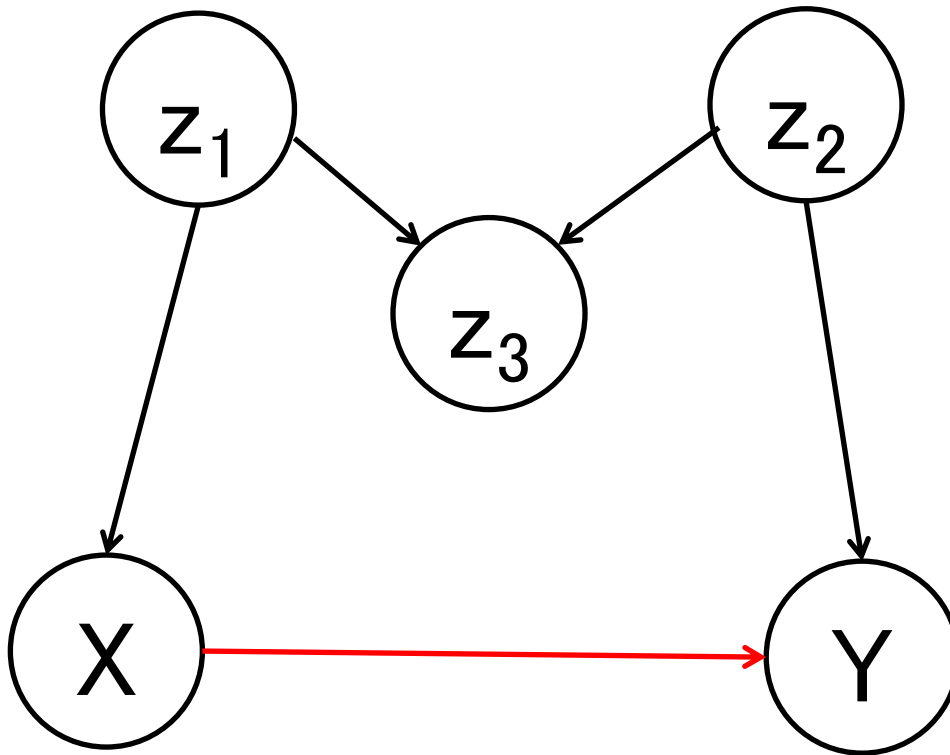
$$(2) Y = b'_0 + b_1 X + b_3 Z_2$$

$$(3) Y = b'_0 + b_1 X + b'_2 Z_1 + b_3 Z_2$$

(1)-(3)いずれの場合でも、理論的には
 X の効果(b_1)は同じ値になる

(参考) : バックドア基準 (M bias)

4変量以上の場合、道(path)で考える



このパターンはMalicious Mとか
M biasと呼ばれる

$X \rightarrow Y$ の効果を正しく推定するためには

(1) $Y = b_1 X$

(2) $Y = b_1 X + b_2 Z_1$

(3) $Y = b_1 X + b_3 Z_2$

(4) $Y = b_1 X + b'_2 Z_1 + b'_3 Z_2$

しかし、(5)はだめ

(5) $Y = b_1^* X + b_4 Z_3$

ただし、(6)はok

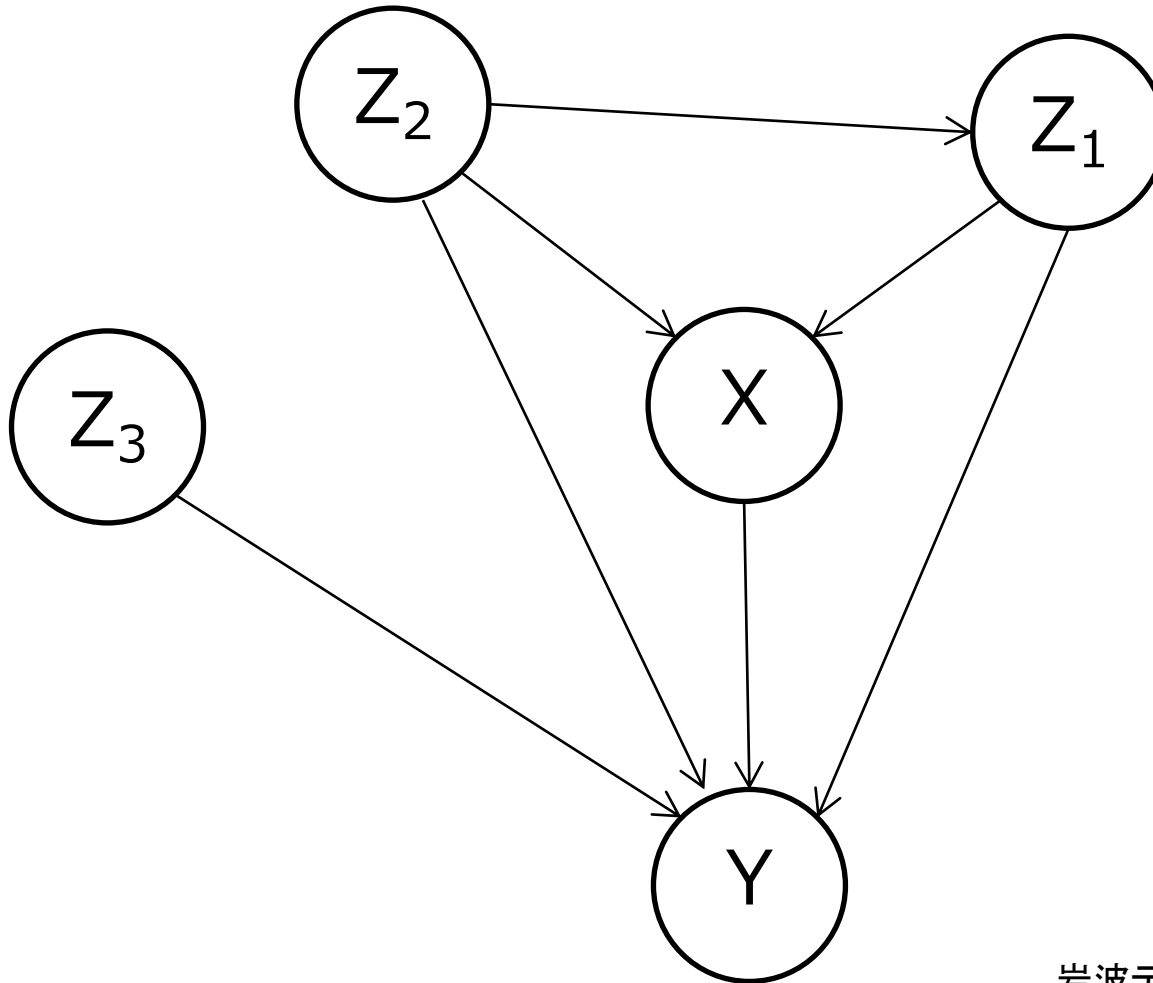
(6) $Y = b_1 X + b''_2 Z_1 + b''_3 Z_2 + b''_4 Z_3$

(1)-(4)(6)のいずれでも、理論的には
 X の効果(b_1)は同じ値になる

(1)-(6)は切片(b_0)⁵⁸省略

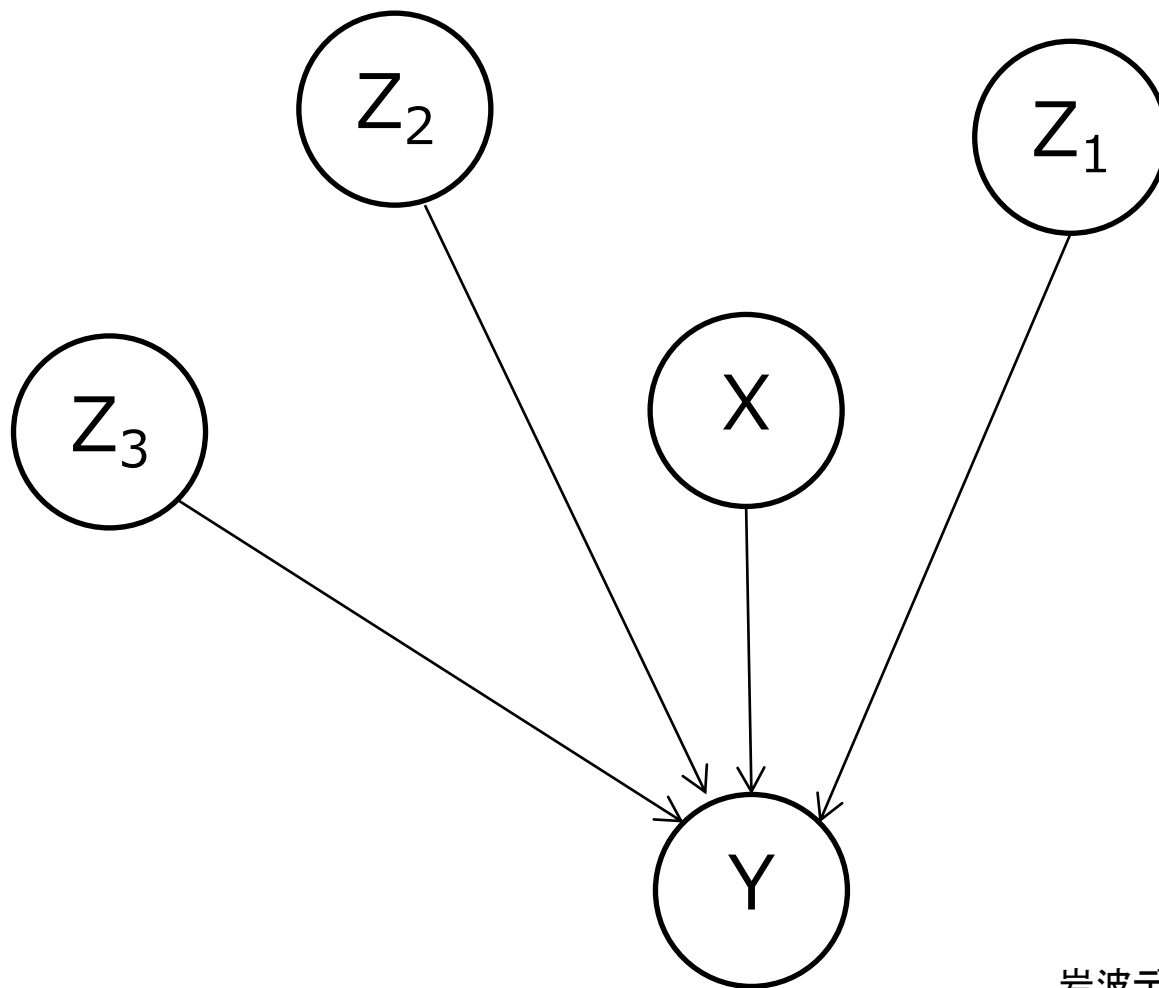
練習問題 (a)

図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？



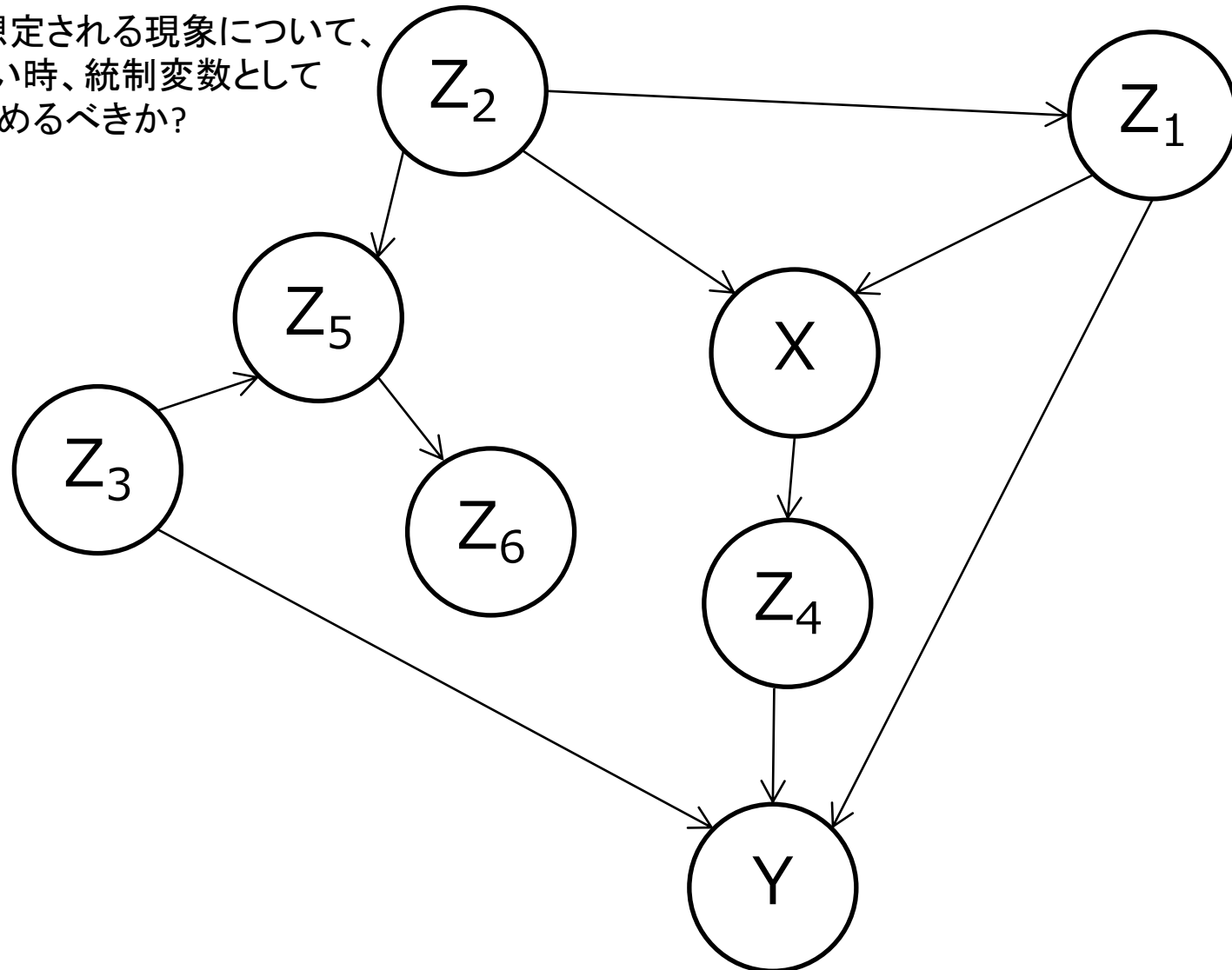
練習問題 (b)

図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？



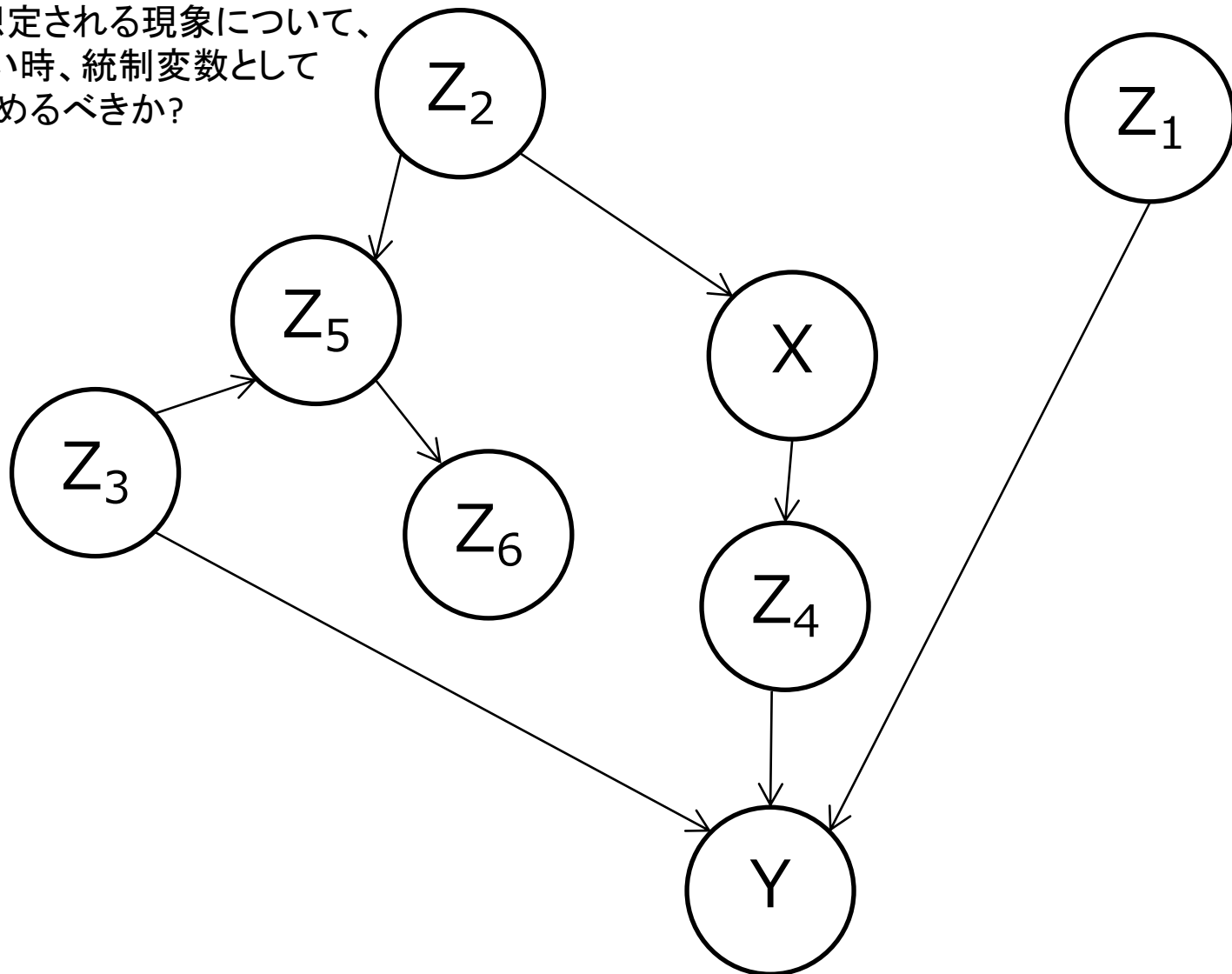
練習問題 (c)

図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？

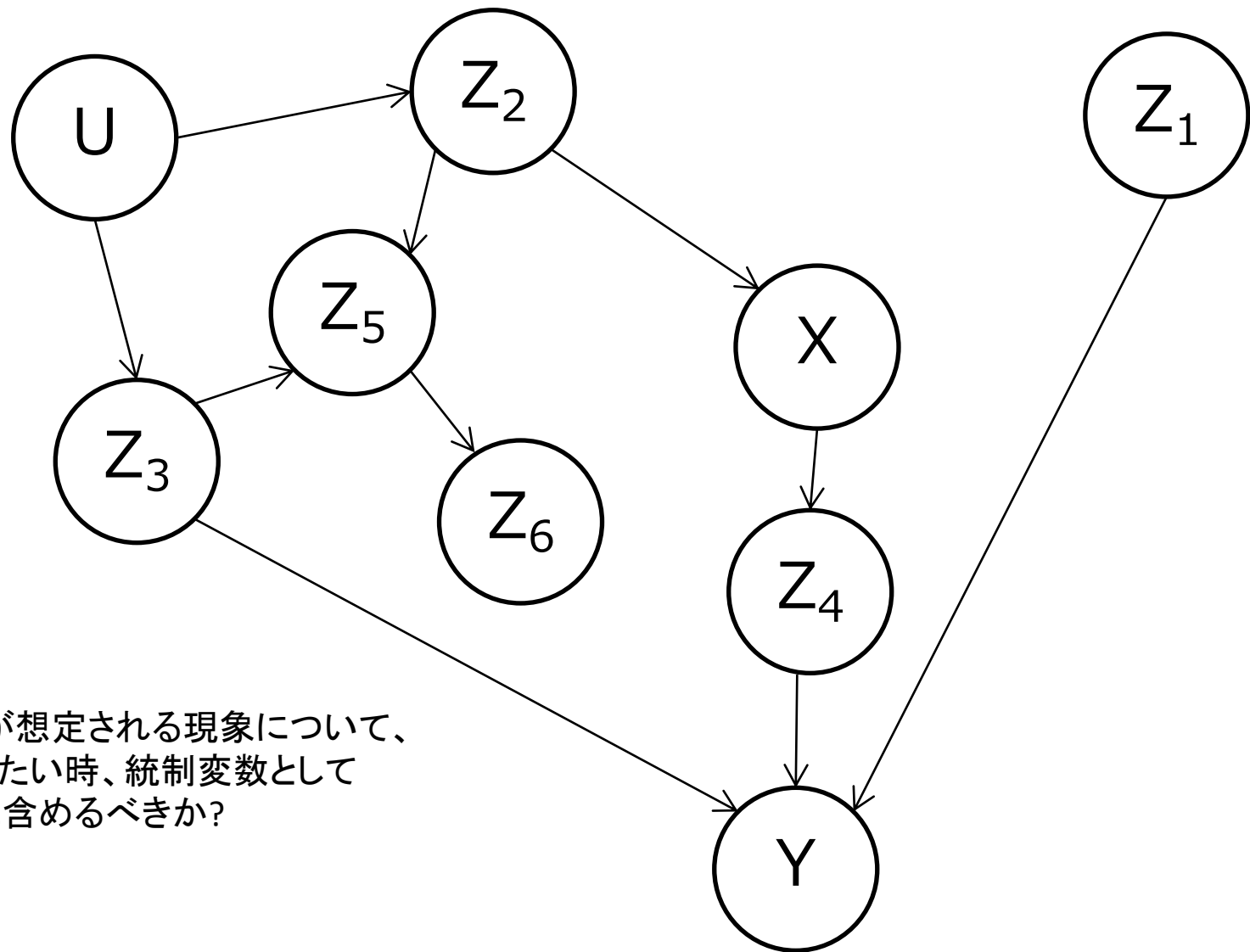


練習問題 (d)

図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？



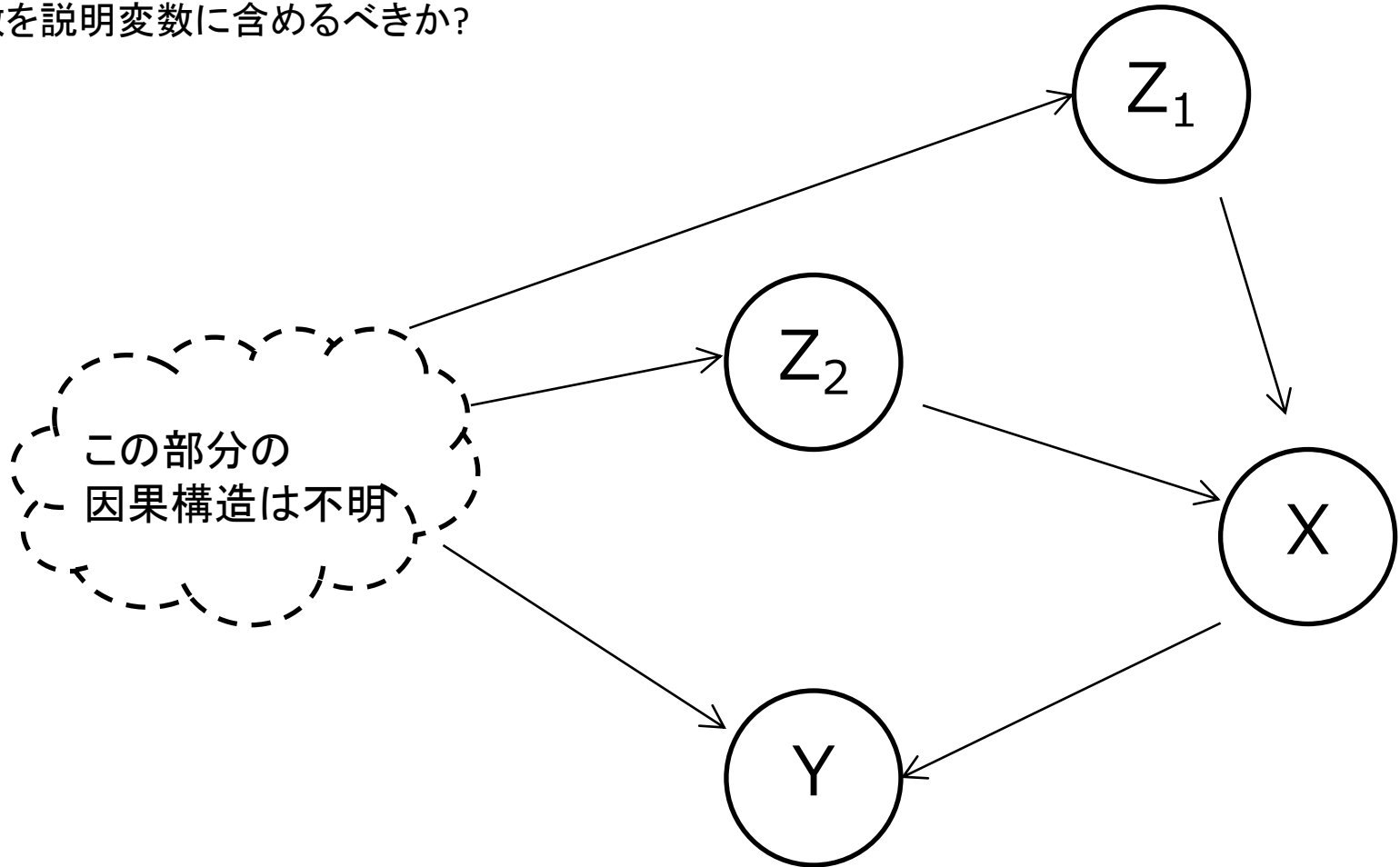
練習問題 (e)



図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？

練習問題 (f)

図のような因果グラフが想定される現象について、
 $X \rightarrow Y$ への効果を推定したい時、統制変数として
どの変数を説明変数に含めるべきか？



解答

正解例

(a) $Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_2 Z_2$

(b) $Y = \beta_0 + \beta_1 X$

(c) $Y = \beta_0 + \beta_1 X + \beta_2 Z_1$
 $Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_2 Z_2$

(d) $Y = \beta_0 + \beta_1 X$
 $Y = \beta_0 + \beta_1 X + \beta_2 Z_2$

(e) $Y = \beta_0 + \beta_1 X + \beta_2 Z_2$
 $Y = \beta_0 + \beta_1 X + \beta_3 Z_3$

(f) $Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_2 Z_2$

別解例・誤答例

$$Y = \beta_0 + \beta_1 X$$

は間違い 理論的には同じ値の β_1 になる

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1$$

も冗長だが正解

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_2 + \beta_5 Z_5$$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_2 + \beta_5 Z_5 + \beta_6 Z_3$$

も冗長だが正解

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_2 + \beta_3 Z_5$$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_2 + \beta_3 Z_5 + \beta_4 Z_3$$

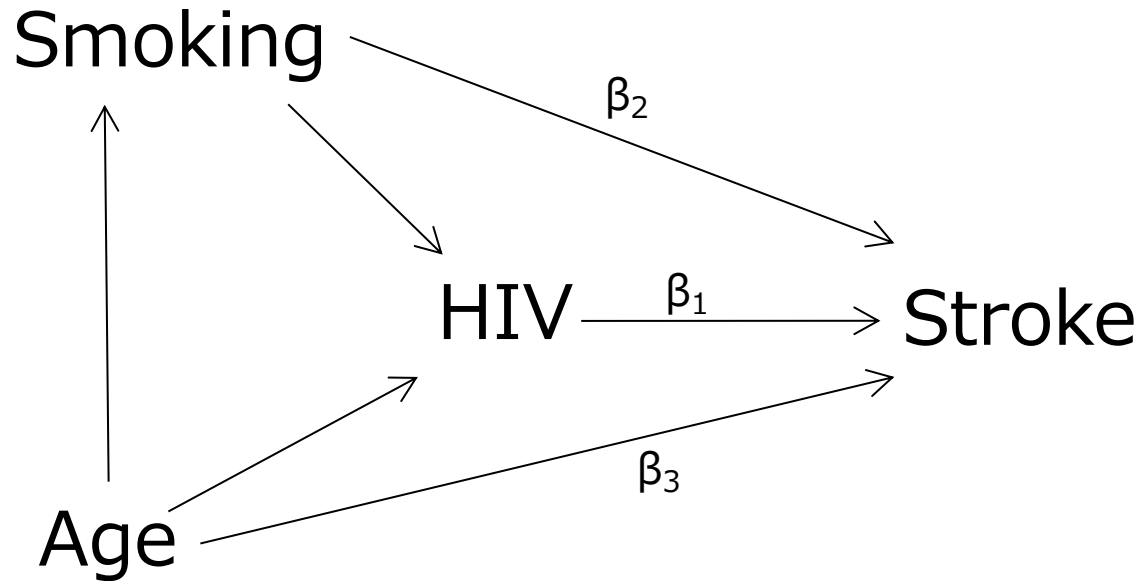
も冗長だが正解

$$Y = \beta_0 + \beta_1 X$$

は間違い

(参考)統制変数の回帰係数は解釈すべきか？

統制変数の偏回帰係数を軽率に解釈することをTable 2 Fallacyとよぶ



上記DAGを念頭に、下記のロジスティック回帰を推定した。

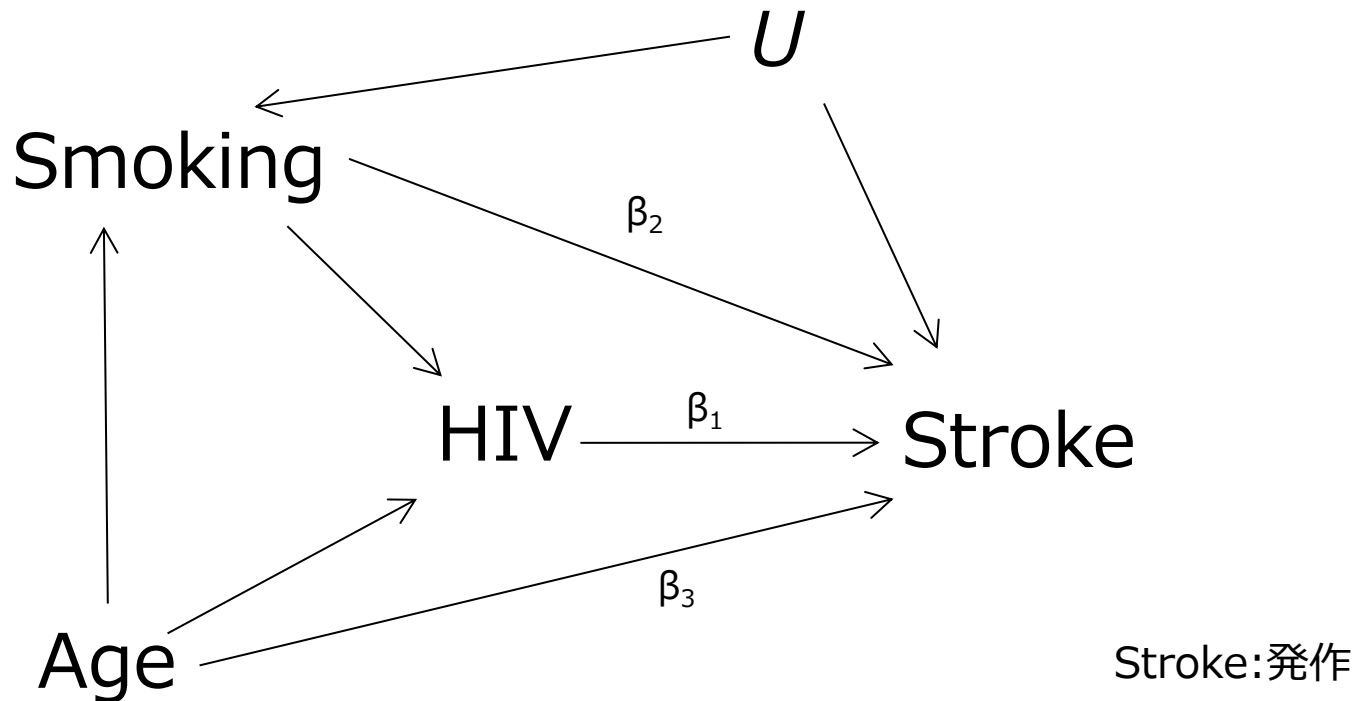
$$\text{logit}(\text{Stroke}) = \beta_0 + \beta_1 \text{HIV} + \beta_2 \text{Smoking} + \beta_3 \text{Age}$$

このとき、 β_1 はHIV→Strokeの効果を正しく示している。
同様に、 β_2 もSmokingの効果を示していると考えて良いか？

(参考)統制変数の回帰係数は解釈すべきか?

$$\text{logit}(\text{Stroke}) = \beta_0 + \beta_1 \times \text{HIV} + \beta_2 \times \text{Smoking} + \beta_3 \times \text{Age}$$

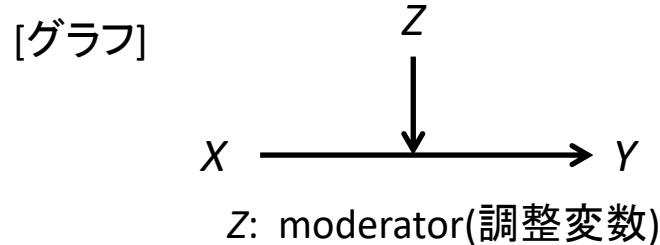
- β_2 を β_1 と同様に解釈してはいけない
(= β_2 はSmoking→Strokeの効果を正しく示していないかもしれない)。
- U という未調整の交絡因子が存在している可能性があるからだ。
- U の存在がない、といえるなら、 β_2 は正しくSmokingの効果を示している



回帰分析の応用

Exercise 8 交互作用モデルと媒介作用モデル

[モデル名] 交互作用モデル XとZは独立している



[モデル] $Y = \beta_1 X + \beta_2 Z + \beta_3 XZ \dots(1)$

[特徴] XのYへの効果はZの値に依存

$$\begin{aligned} Y &= \beta_1 X + \beta_2 Z + \beta_3 XZ \\ &= (\beta_1 + \beta_3 Z)X + \beta_2 Z \end{aligned}$$

[例] ・コンティンジェンシー理論

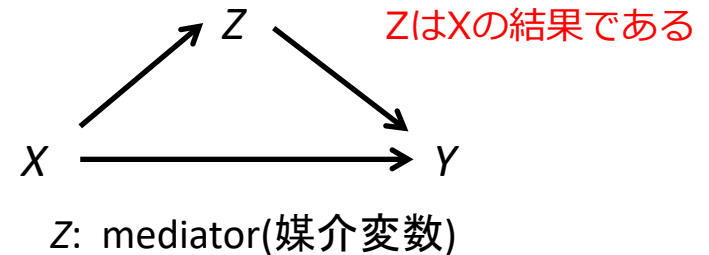
X: 有機的管理, Y: 業績, Z: 不安定環境

[推定法]

(1)式を回帰モデルで推定し β_3 を検討

参考:Holmbeck(1997)

媒介作用モデル(間接効果モデル)



$$Y = \beta_1 X + \beta_2 Z \dots(2)$$

$$Z = \beta_3 X \dots(3)$$

Xは2つのパス経路でYに効果を及ぼす

(i) XがYに影響

(ii) XがZに影響し、さらにZがYに影響

$$Y = \beta_1 X + \beta_2 Z$$

$$= \beta_1 X + \beta_2 \beta_3 X$$

$$= \beta_4 X \dots(4)$$

・有能さのジレンマ

X: 個人能力, Y: 起業率, Z: 大企業就職率

(2)(3)式をSEMで推定し $\beta_1 \beta_2 \beta_3$ を検討

or

(2)(4)式を回帰モデルで推定し $\beta_1 \beta_2 \beta_4$ を検討

交互作用モデル

Exercise 8 交互作用モデル

交互作用モデルは重回帰モデルに積項(交差項, XZ)を追加したもの

典型的には下記のように表す。

$$Y = \beta_1 X + \beta_2 Z + \beta_3 XZ \cdots (1)$$

Z: 調整変数
(moderator)

(1)は

$$\begin{aligned} Y &= \beta_1 X + \beta_2 Z + \beta_3 XZ \\ &= \underline{(\beta_1 + \beta_3 Z)} X + \beta_2 Z \end{aligned}$$

β_1 : 主効果
 β_3 : 交互作用効果

と変形できる。

マージナル効果
(限界効果)

マージナル効果はMEと
略すことあり

交互作用モデルは、Xの効果がZの水準によって変化する

Exercise 8 1.交互作用モデルを推定する

		Dependent variable:		
		mpg		
		(1)	(2)	(3)
主効果ー	wt	-5.344*** (0.559)	-3.878*** (0.633)	-8.217*** (1.270) ③
調整変数	hp		-0.032*** (0.009)	-0.120*** (0.025)
交互作用効果	wt:hp			0.028*** (0.007) ②
統計ソフトRでは wt:hpで wt×hpを表す	Constant	37.285*** (1.878)	37.227*** (1.599)	49.808*** (3.605)
	Observations	32	32	32
	R ²	0.753	0.827	0.885 ①
	Adjusted R ²	0.745	0.815	0.872
	Residual Std. Error	3.046 (df = 30)	2.593 (df = 29)	2.153 (df = 28)
	F Statistic	91.375*** (df = 1; 30)	69.211*** (df = 2; 29)	71.660*** (df = 3; 28)
Note:		*p<0.1; **p<0.05; ***p<0.01		

- ① 決定係数R², 調整済みR²は(2)→(3)で改善している
- ② 交互作用効果は統計的に有意である(交互作用効果あり)
- ③④ 主効果は(2)(3)ともに統計的に有意である。
符号が交互作用と逆転しているので注意

交互作用モデルの留意点

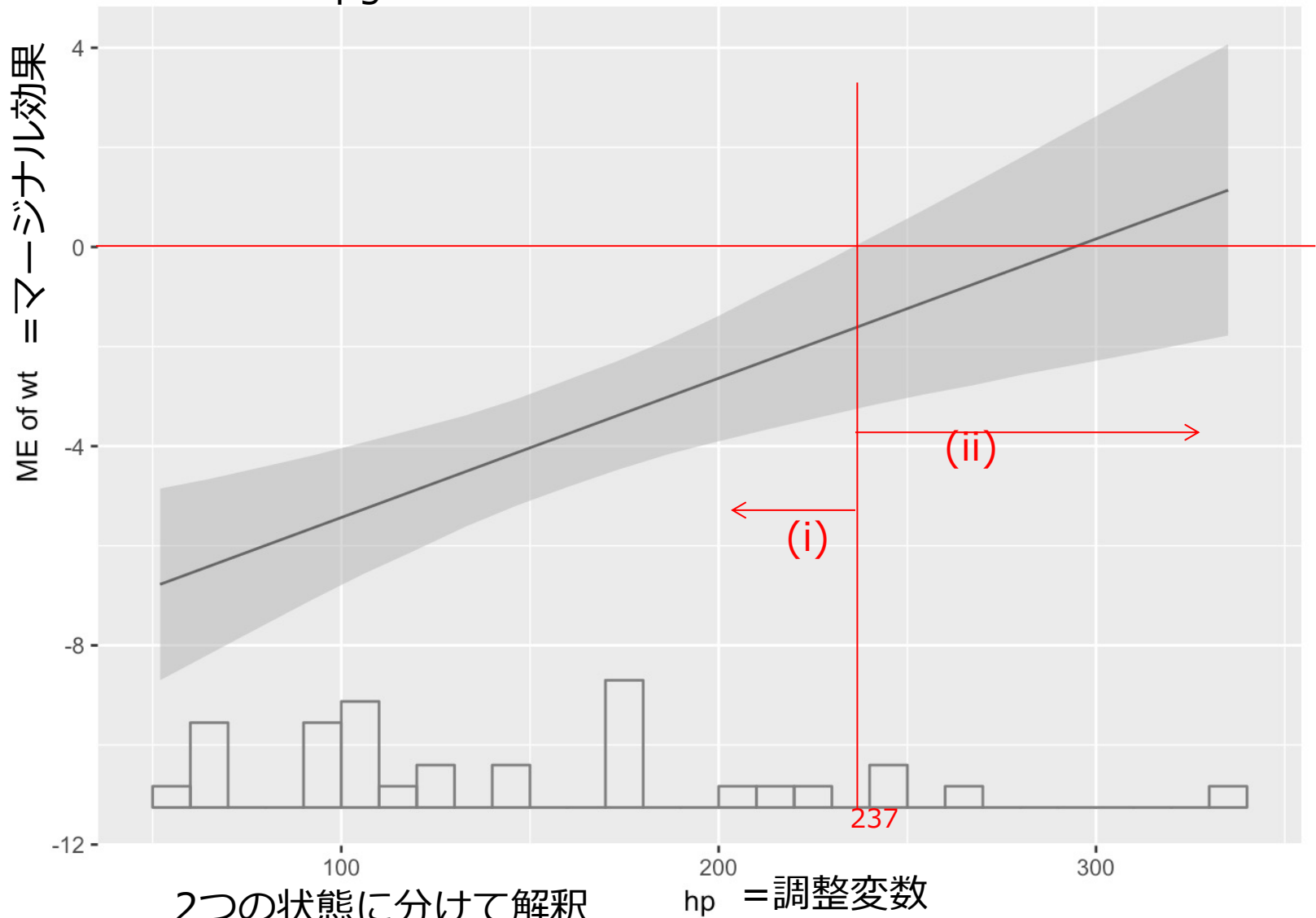
1. 交互作用モデルには交互作用項(XZ)と同時に構成項(XとZ)も含めること

XZで交互作用があるならば、XとZの単純効果もあるはず

2. XZとXは独立して評価することはできない。必ず合わせて評価すること（マージナル効果をみること）
3. Xのマージナル効果はZの水準によって変化することに留意。特にZの取りうる範囲に留意すること。
4. Xのマージナル効果の信頼区間はZの水準によって変化することに留意

Exercise 8 2. マージナル効果図を作図する

mpgに対するwtの効果（マージナル効果）



- (i) $hp < 237$ の場合は、wtの効果は統計的に有意にマイナス
- (ii) $237 < hp$ の場合は、wtの効果は統計的に有意ではない

媒介モデル

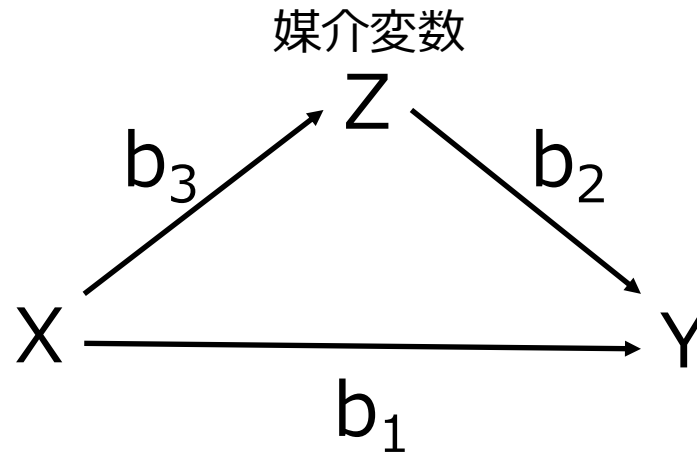
Exercise 9 媒介モデル(構造方程式モデル, SEM)

- 回帰モデルは、**1本の方程式** (=構造方程式) で目的変数と説明変数の関係を表現した
- しかし、1本の方程式ではなく、**複数の方程式**で関係が表現されることもある。
- 複数の方程式で関係を表現したものを、**構造方程式モデル(SEM)**とか、システム方程式モデルと呼ぶ。
- 媒介モデルは、構造方程式モデルの1種

Exercise 9 媒介モデルの例

$$Y = b_1X + b_2Z \quad \cdots(1)$$

$$Z = b_3X \quad \cdots(2)$$



Zを**媒介変数（中間変数）**とよぶ

媒介モデルは、直接効果・間接効果・総合効果が存在する

直接効果：Xから直接的にYに作用する効果(b_1)

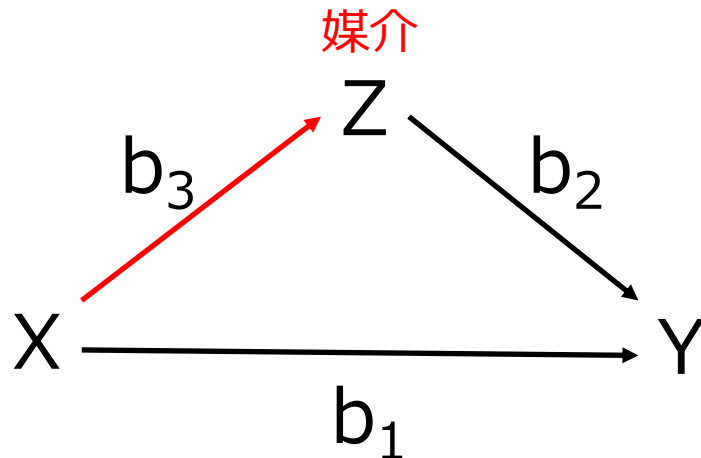
間接効果：XからZを経由してYに作用する効果($b_2 \times b_3$)

総合効果：直接効果と間接効果を合わせたもの($b_1 + b_2 \times b_3$)

Exercise 9 媒介モデルと回帰モデルとの比較

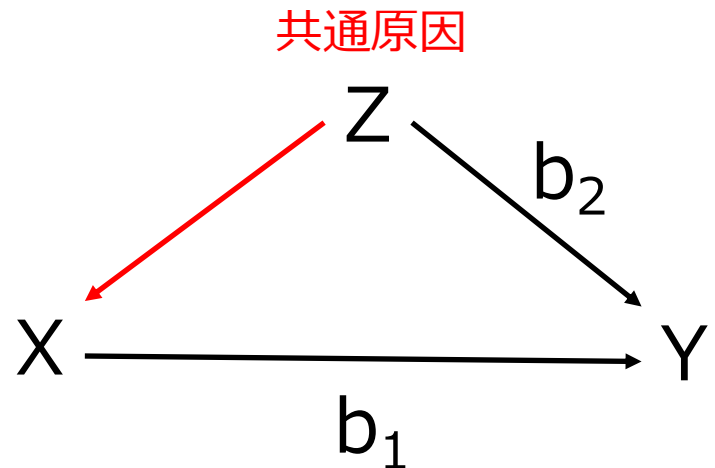
$$Y = b_1X + b_2Z \quad \cdots(1)$$

$$Z = b_3X \quad \cdots(2)$$



媒介モデル

$$Y = b_1X + b_2Z \quad \cdots(1)$$



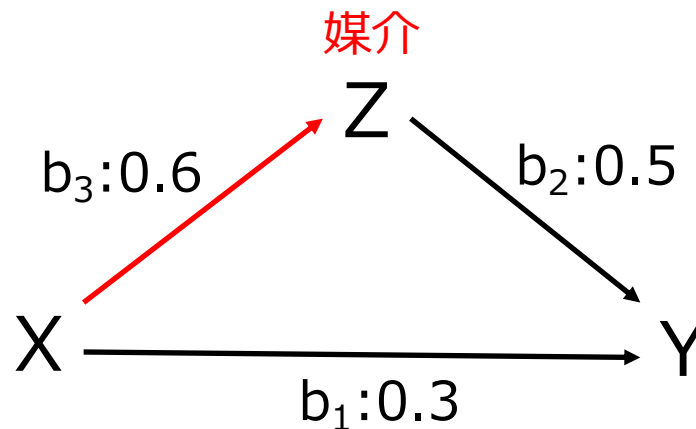
通常の重回帰モデル

Exercise 9 数値例(1)

モデル式

$$y = 0.3x + 0.5z$$

$$z = 0.6x$$



設定したパラメータ

$x \rightarrow y$ 直接効果: 0.3

$x \rightarrow y$ 間接効果: $0.6 * 0.5 = 0.3$

$x \rightarrow y$ 総合効果: $0.3 + 0.6 * 0.5 = 0.6$

Exercise 9 数値例(1) 重回帰モデル

Dependent variable:		
	y	
	(1)	(2)
x	0.581*** (0.033)	0.301*** (0.036)
z		0.437*** (0.032)
Constant	0.029 (0.034)	0.025 (0.031)
Observations	1,000	1,000
R2	0.236	0.358
Adjusted R2	0.235	0.357
Residual Std. Error	1.068 (df = 998)	0.979 (df = 997)
F Statistic	308.370*** (df = 1; 998)	278.312*** (df = 2; 997)

Note:

*p<0.1; **p<0.05; ***p<0.01

(参考) Exercise 9 数値例(1) 重回帰モデル

サンプルサイズが小さい時は、統計的有意性あり→統計的有意なし、になる

Dependent variable:			
	y		
	(1)		(2)
x	統計的有意性あり 0.525*** (0.161)	→ 効果が小さくなる	0.229 (0.150) 統計的有意性なし
z			0.608*** (0.133)
Constant	0.134 (0.156)		0.038 (0.133)
Observations	50		50 サンプルサイズが小さい
R2	0.181		0.432
Adjusted R2	0.164		0.408
Residual Std. Error	1.075 (df = 48)		0.904 (df = 47)
F Statistic	10.618*** (df = 1; 48)		17.894*** (df = 2; 47)
Note: *p<0.1; **p<0.05; ***p<0.01			

Exercise 9 数値例(1) SEMの推定のやり方

```
library(lavaan)
```

} SEMのためのパッケージ読込

```
model_00 = "
```

```
y ~ b1*x + b2*z
```

```
z ~ b3*x
```

```
ind := b2*b3
```

```
total := b1 + (b2*b3)
```

```
"
```

} SEMのためのモデル式記述

SEM推定

```
fit_sem_00 = sem(model_00, data=test_data_00)
```

推定結果の表示

```
summary(fit_sem_00, fit.measures=TRUE)
```

Exercise 9 数値例(1) SEMの推定結果テーブル

lavaan 0.6-5 ended normally after 11 iterations

Parameter Estimates:

Estimator ML
Optimization method NLMINB
Number of free parameters 5

Information
Information saturated (h1) model
Standard errors
Expected
Structured
Standard

Number of observations 1000

Regressions:

Model Test User Model:

	Estimate	Std.Err	z-value	P(> z)
y ~ x	(b1) 0.301	0.036	8.270	0.000
z ~ x	(b2) 0.437	0.032	13.801	0.000
z ~ x	(b3) 0.639	0.030	21.149	0.000

Test statistic 0.000
Degrees of freedom 0

Model Test Baseline Model:

Variances:

Test statistic 813.275
Degrees of freedom 3
P-value 0.000

	Estimate	Std.Err	z-value	P(> z)
.y	0.956	0.043	22.361	0.000
.z	0.953	0.043	22.361	0.000

User Model versus Baseline Model:

Defined Parameters:

Comparative Fit Index (CFI) 1.000
Tucker-Lewis Index (TLI) 1.000

CFI > 0.95が目安

	Estimate	Std.Err	z-value	P(> z)
ind	0.280	0.024	11.558	0.000
total	0.581	0.033	17.578	0.000

Loglikelihood and Information Criteria:

Loglikelihood user model (H0) -2790.937
Loglikelihood unrestricted model (H1) -2790.937

Akaike (AIC) 5591.874
Bayesian (BIC) 5616.413
Sample-size adjusted Bayesian (BIC) 5600.532

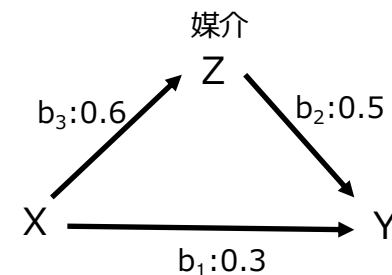
Root Mean Square Error of Approximation:

RSMEA < 0.05が目安

RMSEA 0.000
90 Percent confidence interval - lower 0.000
90 Percent confidence interval - upper 0.000
P-value RMSEA <= 0.05 NA

Standardized Root Mean Square Residual:

SRMR 0.000



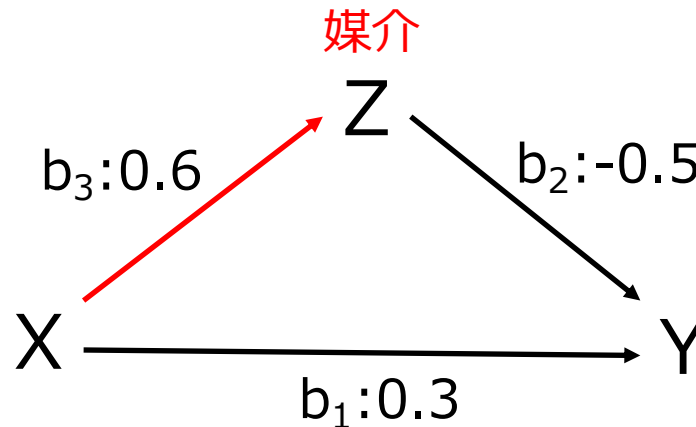
モデルの適合度の指標
(今回は説明省略)

Exercise 9 数値例(2)

モデル式

$$y = 0.3x + (-0.5)z$$

$$z = 0.6x$$



$x \rightarrow y$ 直接効果: 0.3

$x \rightarrow y$ 間接効果: $0.6 * (-0.5) = -0.3$

$x \rightarrow y$ 総合効果: $0.3 + 0.6 * 0.5 = 0.0$

直接効果はプラス

間接効果はマイナス

Exercise 9 数值例(2) 重回帰モデル

Dependent variable:

y

	(1)	(2)
x	総合効果 0.013 (0.038)	0.337*** 直接効果 (0.037)
z		-0.582*** (0.032)
Constant	0.052 (0.037)	0.048 (0.032)
Observations	1,000	1,000
R2	0.0001	0.248
Adjusted R2	-0.001	0.246
Residual Std. Error	1.162 (df = 998)	1.008 (df = 997)
F Statistic	0.112 (df = 1; 998)	164.324*** (df = 2; 997)

Note:

*p<0.1; **p<0.05; ***p<0.01

Exercise 9 数値例(2) SEMの推定のやり方

```
library(lavaan)
```

} SEMのためのパッケージ読込

```
model_01 = "
```

```
y ~ b1*x + b2*z
```

```
z ~ b3*x
```

```
ind := b2*b3
```

```
total := b1 + (b2*b3)
```

```
"
```

} SEMのためのモデル式記述

SEM推定

```
fit_sem_01 = sem(model_01, data=test_data_01)
```

推定結果の表示

```
summary(fit_sem_01, fit.measure=TRUE)
```

Exercise 9 数値例(2) SEMの推定結果テーブル

lavaan 0.6-5 ended normally after 11 iterations

Estimator	ML
Optimization method	NLMINB
Number of free parameters	5
Number of observations	1000

Model Test User Model:

Test statistic	0.000
Degrees of freedom	0

Model Test Baseline Model:

Test statistic	545.526
Degrees of freedom	3
P-value	0.000

User Model versus Baseline Model:

<u>Comparative Fit Index (CFI)</u>	<u>1.000</u>
Tucker-Lewis Index (TLI)	1.000

Loglikelihood and Information Criteria:

Loglikelihood user model (H0)	-2838.466
Loglikelihood unrestricted model (H1)	-2838.466
Akaike (AIC)	5686.932
Bayesian (BIC)	5711.471
Sample-size adjusted Bayesian (BIC)	5695.590

Root Mean Square Error of Approximation:

<u>RMSEA</u>	<u>0.000</u>
90 Percent confidence interval - lower	0.000
90 Percent confidence interval - upper	0.000
P-value RMSEA <= 0.05	NA

Standardized Root Mean Square Residual:

SRMR	0.000
------	-------

Parameter Estimates:

Information	Expected
Information saturated (h1) model	Structured
Standard errors	Standard

Regressions:

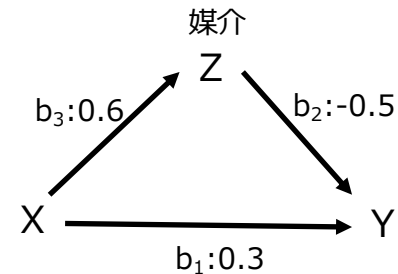
	Estimate	Std.Err	z-value	P(> z)
y ~				
x	<u>(b1) 0.337</u>	0.037	9.034	0.000
z	<u>(b2) -0.582</u>	0.032	-18.152	0.000
z ~				
x	<u>(b3) 0.557</u>	0.032	17.255	0.000

Variances:

	Estimate	Std.Err	z-value	P(> z)
.y	1.014	0.045	22.361	0.000
.z	0.988	0.044	22.361	0.000

Defined Parameters:

	Estimate	Std.Err	z-value	P(> z)
ind	<u>-0.324</u>	0.026	-12.506	0.000
total	<u>0.013</u>	0.038	0.335	0.738



SEMなら正しいモデルを選択できるか？

残念ながら、

SEMだから正しいモデルを選択できる、
ということは無い

正しくないモデルであっても、

- ・ 適合度の目安をクリアする
- ・ 複数の異なるモデルが適合度をクリアする

適合度の目安クリアは、

「仮説モデルであったとしても、データと矛盾しない」
といっているだけ。

Exercise 9 数値例(2) SEMの推定結果テーブル

lavaan 0.6-5 ended normally after 11 iterations

Parameter Estimates:

Estimator	ML
Optimization method	NLMINB
Number of free parameters	3

Information	Expected
Information saturated (h1) model	Structured
Standard errors	Standard

Number of observations	1000
------------------------	------

Model Test User Model:

Test statistic	0.000
Degrees of freedom	0

Regressions:

		Estimate	Std.Err	z-value	P(> z)
y ~					
x	(b1)	0.337	0.037	9.034	0.000
z	(b2)	-0.582	0.032	-18.152	0.000

Model Test Baseline Model:

Test statistic	284.906
Degrees of freedom	2
P-value	0.000

Variances:

	Estimate	Std.Err	z-value	P(> z)
.y	1.014	0.045	22.361	0.000

User Model versus Baseline Model:

Comparative Fit Index (CFI)	1.000
Tucker-Lewis Index (TLI)	1.000

CFI > 0.95が目安

Loglikelihood and Information Criteria:

Loglikelihood user model (H0)	-1425.764
Loglikelihood unrestricted model (H1)	-1425.764

Akaike (AIC)	2857.528
Bayesian (BIC)	2872.251
Sample-size adjusted Bayesian (BIC)	2862.723

Root Mean Square Error of Approximation: RSMEA < 0.05が目安

RMSEA	0.000
90 Percent confidence interval - lower	0.000
90 Percent confidence interval - upper	0.000
P-value RMSEA <= 0.05	NA

Standardized Root Mean Square Residual:

SRMR	0.000
------	-------

分析モデル

$$y = b_1x + b_2z$$

データ生成に使ったモデル

$$y = 0.3x + (-0.5)z$$

$$z = 0.6x$$

分析モデルは、データ生成したモデルとは異なるが、適合度は目安を合格している。

まとめ：この講義でやったこと

- 統計分析を行う環境として、Rを準備した
- Rを使ってデータを操作し、初歩的な統計分析をした
- 重回帰分析と、その分析結果の解釈
- 交互作用を含んだ回帰モデル
- 媒介モデル（構造方程式モデル）
- 統制変数（交絡変数）とバックドア基準